VOLUME 3,NUMBER 1 JANUARY 2005

ISSN:1548-5390 PRINT,1559-176X ONLINE



JOURNAL

OF CONCRETE

AND APPLICABLE

MATHEMATICS

EUDOXUS PRESS,LLC

SCOPE AND PRICES OF THE JOURNAL
Journal of Concrete and Applicable Mathematics

A quartely international publication of **Eudoxus Press,LLC**

Editor in Chief: George Anastassiou
Department of Mathematical Sciences,
University of Memphis
Memphis, TN 38152, U.S.A.
ganastss@memphis.edu

The main purpose of the "Journal of Concrete and Applicable Mathematics" is to publish high quality original research articles from all subareas of Non-Pure and/or Applicable Mathematics and its many real life applications, as well connections to other areas of Mathematical Sciences, as long as they are presented in a Concrete way. It welcomes also related research survey articles and book reviews.A sample list of connected mathematical areas with this publication includes and is not restricted to: Applied Analysis, Applied Functional Analysis, Probability theory, Stochastic Processes, Approximation Theory, O.D.E, P.D.E, Wavelet, Neural Networks,Difference Equations, Summability, Fractals, Special Functions, Splines, Asymptotic Analysis, Fractional Analysis, Inequalities, Moment Theory, Numerical Functional Analysis,Tomography, Asymptotic Expansions, Fourier Analysis, Applied Harmonic Analysis, Integral Equations, Signal Analysis, Numerical Analysis, Optimization, Operations Research, Linear Programming, Fuzzyness, Mathematical Finance, Stochastic Analysis, Game Theory, Math.Physics aspects, Applied Real and Complex Analysis, Computational Number Theory, Graph Theory, Combinatorics,
Computer Science Math.related topics,combinations of the above, etc. In general any kind of Concretely presented Mathematics which is Applicable fits to the scope of this journal.
Working Concretely and in Applicable Mathematics has become a main trend in many recent years,so we can understand better and deeper and solve the important problems of our real and scientific world.
"Journal of Concrete and Applicable Mathematics" is a peer- reviewed International Quarterly Journal.
We are calling for papers for possible publication. The contributor should send three copies of the contribution to the editor in-Chief typed in TEX, LATEX double spaced. [See: Instructions to Contributors]

Journal of Concrete and Applicable Mathematics(JCAAM)
ISSN:1548-5390 PRINT, 1559-176X ONLINE.
is published in January,April,July and October of each year by
EUDOXUS PRESS,LLC,
1424 Beaver Trail Drive,Cordova,TN38016,USA,
Tel.001-901-751-3553
anastassioug@yahoo.com
http://www.EudoxusPress.com.

Annual Subscription Current Prices:For USA and Canada,Institutional:Print $250,Electronic $220,Print and Electronic $310.Individual:Print $77,Electronic $60,Print &Electronic $110.For any other part of the world add $25 more to the above prices for Print.

Single article PDF file for individual $8.Single issue in PDF form for individual $25.
The journal carries page charges $8 per page of the pdf file of an article,payable upon acceptance of the article within one month and before publication.
No credit card payments.Only certified check,money order or international check in US dollars are acceptable.
Combination orders of any two from JoCAAA,JCAAM,JAFA receive 25% discount,all three receive 30% discount.

Copyright©2004 by Eudoxus Press,LLC all rights reserved.JCAAM is printed in USA.

JCAAM is reviewed and abstracted by AMS Mathematical Reviews,MATHSCI,and Zentralblaat MATH.

It is strictly prohibited the reproduction and transmission of any part of JCAAM and in any form and by any means without the written permission of the publisher.It is only allowed to educators to Xerox articles for educational purposes.The publisher assumes no responsibility for the content of published papers.
JCAAM IS A JOURNAL OF RAPID PUBLICATION

Editorial Board
Associate Editors

Editor in -Chief:
George Anastassiou
Department of Mathematical Sciences
The University Of Memphis
Memphis,TN 38152,USA
tel.901-678-3144,fax 901-678-2480
e-mail ganastss@memphis.edu
www.msci.memphis.edu/~ganastss/jcaam
Areas:Approximation Theory,
Probability,Moments,Wavelet,
Neural Networks,Inequalities,Fuzzyness.

Associate Editors:

1) Ravi Agarwal
Florida Institute of Technology
Applied Mathematics Program
150 W.University Blvd.
Melbourne,FL 32901,USA
agarwal@fit.edu
Differential Equations,Difference
Equations,inequalities

2) Shair Ahmad
University of Texas at San Antonio
Division of Math.& Stat.
San Antonio,TX 78249-0664,USA
SAHMAD@utsa.edu
Differential Equations,Mathematical
Biology

3) Drumi D.Bainov
Medical University of Sofia
P.O.Box 45,1504 Sofia,Bulgaria
drumibainov@yahoo.com
Differential Equations,Optimal Control,
Numerical Analysis,Approximation Theory

4) Carlo Bardaro
Dipartimento di Matematica & Informatica
Universita' di Perugia
Via Vanvitelli 1
06123 Perugia,ITALY
tel.+390755855034, +390755853822,
fax +390755855024
bardaro@unipg.it ,
bardaro@dipmat.unipg.it
Functional Analysis and Approximation Th.,

19) Rupert Lasser
Institut fur Biomathematik & Biomertie,GSF
-National Research Center for environment and
health
Ingolstaedter landstr.1
D-85764 Neuherberg,Germany
lasser@gsf.de
Orthogonal Polynomials,Fourier Analysis,
Mathematical Biology

20) Alexandru Lupas
University of Sibiu
Faculty of Sciences
Department of Mathematics
Str.I.Ratiu nr.7
2400-Sibiu,Romania
lupas@ulbsibiu.ro
Classical Analysis,Inequalities,
Special Functions,Umbral Calculus,
Approximation Th.,Numerical Analysis
and Methods

21) Ram N.Mohapatra
Department of Mathematics
University of Central Florida
Orlando,FL 32816-1364
tel.407-823-5080
ramm@mail.ucf.edu
Real and Complex analysis,Approximation Th.,
Fourier Analysis, Fuzzy Sets and Systems

22) Rainer Nagel
Arbeitsbereich Funktionalanalysis
Mathematisches Institut
Auf der Morgenstelle 10
D-72076 Tuebingen
Germany
tel.49-7071-2973242
fax 49-7071-294322
rana@fa.uni-tuebingen.de
evolution equations,semigroups,spectral th.,
positivity

23) Panos M.Pardalos
Center for Appl. Optimization
University of Florida
303 Weil Hall
P.O.Box 116595
Gainesville,FL 32611-6595

Summability,Signal Analysis,Integral Equations,
Measure Th.,Real Analysis

5) Francoise Bastin
Institute of Mathematics
University of Liege
4000 Liege
BELGIUM
f.bastin@ulg.ac.be
Functional Analysis,Wavelets

6) Paul L.Butzer
RWTH Aachen
Lehrstuhl A fur Mathematik
D-52056 Aachen
Germany
tel.0049/241/80-94627 office,
0049/241/72833 home,
fax 0049/241/80-92212
Butzer@rwth-aachen.de
Approximation Th.,Sampling Th.,Signals,
Semigroups of Operators,Fourier Analysis

7) Yeol Je Cho
Department of Mathematics Education
College of Education
Gyeongsang National University
Chinju 660-701
KOREA
tel.055-751-5673 Office,
055-755-3644 home,
fax 055-751-6117
yjcho@nongae.gsnu.ac.kr
Nonlinear operator Th.,Inequalities,
Geometry of Banach Spaces

8) Sever S.Dragomir
School of Communications and Informatics
Victoria University of Technology
PO Box 14428
Melbourne City M.C
Victoria 8001,Australia
tel 61 3 9688 4437,fax 61 3 9688 4050
sever.dragomir@vu.edu.au,
sever@sci.vu.edu.au
Math.Analysis,Inequalities,Approximation Th.,
Numerical Analysis, Geometry of Banach Spaces,
Information Th. and Coding

9) A.M.Fink
Department of Mathematics
Iowa State University
Ames,IA 50011-0001,USA
tel.352-392-9011
pardalos@ufl.edu
Optimization,Operations Research

24) Svetlozar T.Rachev
Dept.of Statistics and Applied Probability Program
University of California,Santa Barbara
CA 93106-3110,USA
tel.805-893-4869
rachev@pstat.ucsb.edu
AND
Chair of Econometrics and Statistics
School of Economics and Business Engineering
University of Karlsruhe
Kollegium am Schloss,Bau II,20.12,R210
Postfach 6980,D-76128,Karlsruhe,Germany
tel.011-49-721-608-7535
rachev@lsoe.uni-karlsruhe.de
Mathematical and Empirical Finance,
Applied Probability, Statistics and Econometrics

25) Paolo Emilio Ricci
Universita' degli Studi di Roma "La Sapienza"
Dipartimento di Matematica-Istituto "G.Castelnuovo"
P.le A.Moro,2-00185 Roma,ITALY
tel.++39 0649913201,fax ++39 0644701007
riccip@uniroma1.it,Paoloemilio.Ricci@uniroma1.it
Orthogonal Polynomials and Special functions,
Numerical Analysis, Transforms,Operational Calculus,
Differential and Difference equations

26) Cecil C.Rousseau
Department of Mathematical Sciences
The University of Memphis
Memphis,TN 38152,USA
tel.901-678-2490,fax 901-678-2480
ccrousse@memphis.edu
Combinatorics,Graph Th.,
Asymptotic Approximations,
Applications to Physics

27) Tomasz Rychlik
Institute of Mathematics
Polish Academy of Sciences
Chopina 12,87100 Torun, Poland
T.Rychlik@impan.gov.pl
Mathematical Statistics,Probabilistic Inequalities

28) Bl. Sendov
Institute of Mathematics and Informatics
Bulgarian Academy of Sciences
Sofia 1090,Bulgaria

tel.515-294-8150
fink@math.iastate.edu
Inequalities,Ordinary Differential Equations

10) Sorin Gal
Department of Mathematics
University of Oradea
Str.Armatei Romane 5
3700 Oradea,Romania
galso@uoradea.ro
Approximation Th.,Fuzzyness,Complex Analysis

11) Jerome A.Goldstein
Department of Mathematical Sciences
The University of Memphis,
Memphis,TN 38152,USA
tel.901-678-2484
jgoldste@memphis.edu
Partial Differential Equations,
Semigroups of Operators

12) Heiner H.Gonska
Department of Mathematics
University of Duisburg
Duisburg,D-47048
Germany
tel.0049-203-379-3542 office
gonska@informatik.uni-duisburg.de
Approximation Th.,Computer Aided Geometric Design

13) Dmitry Khavinson
Department of Mathematical Sciences
University of Arkansas
Fayetteville,AR 72701,USA
tel.(479)575-6331,fax(479)575-8630
dmitry@uark.edu
Potential Th.,Complex Analysis,Holomorphic PDE,Approximation Th.,Function Th.

14) Virginia S.Kiryakova
Institute of Mathematics and Informatics
Bulgarian Academy of Sciences
Sofia 1090,Bulgaria
virginia@diogenes.bg
Special Functions,Integral Transforms,
Fractional Calculus

15) Hans-Bernd Knoop
Institute of Mathematics
Gerhard Mercator University
D-47048 Duisburg
Germany
tel.0049-203-379-2676
bsendov@bas.bg
Approximation Th.,Geometry of Polynomials,
Image Compression

29) Igor Shevchuk
Faculty of Mathematics and Mechanics
National Taras Shevchenko
University of Kyiv
252017 Kyiv
UKRAINE
shevchuk@univ.kiev.ua
Approximation Theory

30) H.M.Srivastava
Department of Mathematics and Statistics
University of Victoria
Victoria,British Columbia V8W 3P4
Canada
tel.250-721-7455 office,250-477-6960 home,
fax 250-721-8962
harimsri@math.uvic.ca
Real and Complex Analysis,Fractional Calculus and Appl.,
Integral Equations and Transforms,Higher Transcendental
Functions and Appl.,q-Series and q-Polynomials,
Analytic Number Th.

31) Ferenc Szidarovszky
Dept.Systems and Industrial Engineering
The University of Arizona
Engineering Building,111
PO.Box 210020
Tucson,AZ 85721-0020,USA
szidar@sie.arizona.edu
Numerical Methods,Game Th.,Dynamic Systems,
Multicriteria Decision making,
Conflict Resolution,Applications
in Economics and Natural Resources Management

32) Gancho Tachev
Dept.of Mathematics
Univ.of Architecture,Civil Eng. and Geodesy
1 Hr.Smirnenski blvd
BG-1421 Sofia,Bulgaria
Approximation Theory

33) Manfred Tasche
Department of Mathematics
University of Rostock
D-18051 Rostock
Germany
manfred.tasche@mathematik.uni-rostock.de
Approximation Th.,Wavelet,Fourier Analysis,
Numerical Methods,Signal Processing,

knoop@math.uni-duisburg.de
Approximation Theory,Interpolation

16) Jerry Koliha
Dept. of Mathematics & Statistics
University of Melbourne
VIC 3010,Melbourne
Australia
koliha@unimelb.edu.au
Inequalities,Operator Theory,
Matrix Analysis,Generalized Inverses

17) Mustafa Kulenovic
Department of Mathematics
University of Rhode Island
Kingston,RI 02881,USA
kulenm@math.uri.edu
Differential and Difference Equations

18) Gerassimos Ladas
Department of Mathematics
University of Rhode Island
Kingston,RI 02881,USA
gladas@math.uri.edu
Differential and Difference Equations
Image Processing,Harmonic Analysis

34) Chris P.Tsokos
Department of Mathematics
University of South Florida
4202 E.Fowler Ave.,PHY 114
Tampa,FL 33620-5700,USA
profcpt@math.usf.edu,profcpt@chuma1.cas.usf.edu
Stochastic Systems,Biomathematics,
Environmental Systems,Reliability Th.

35) Lutz Volkmann
Lehrstuhl II fuer Mathematik
RWTH-Aachen
Templergraben 55
D-52062 Aachen
Germany
volkm@math2.rwth-aachen.de
Complex Analysis,Combinatorics,Graph Theory

JOURNAL OF CONCRETE AND APPLICABLE MATHEMATICS,VOL.3,NO.1,9-29,2005,COPYRIGHT 2005 EUDOXUS PRESS,LLC

Hermite-Kampé de Fériet polynomials and solutions of Boundary Value Problems in the half-space

Giulia Maroscia

Università di Roma "La Sapienza", Dipartimento di Metodi e Modelli Matematici per le Scienze Applicate, Via A.Scarpa, 16, 00161 Roma, Italia
e-mail: gmaroscia@dmmm.uniroma1.it

Paolo E. Ricci

Università di Roma "La Sapienza", Dipartimento di Matematica, P.le A. Moro, 2 00185 Roma, Italia - e-mail: riccip@uniroma1.it

Abstract

Explicit solutions of multidimensional canonical BVP of parabolic, hyperbolic and elliptic type in the half-space are derived by using operational methods and series expansions in terms of the Hermite-Kampé de Fériet polynomials.

2000 Mathematics Subject Classification. 33C45, 44A45, 35G15.
Key words and phrases. Hermite-Kampé de Fériet (or Gould-Hopper) polynomials, Operational calculus, Boundary value problems.

1 Introduction

In a preceding article [5], the two-dimensional polynomials considered by Hermite, and subsequently studied by P. Appell and J. Kampé de Fériet [1], H.W. Gould and A.T. Hopper [12], (see also: [16], p. 76, Eq. 1.9 (6)), G. Dattoli et al. [7], were stressed in order to obtain explicit solutions of all the canonical (parabolic, hyperbolic or elliptic) BVP in the half-plane.

For shortness, in the following, we will use the abbreviation H-KdF to denote the Hermite-Kampé de Fériet polynomials.

In the present article, by using the same operational approach recalled in [6], we present here some generalizations of the results obtained in [5] to the case of the solution of all the canonical BVP in the half-space.

According to our results, the two-dimensional H-KdF polynomials appear as the natural tool for representing the solutions of all these problems.

It is worth noting that the use of operational methods in this framework naturally leads to the consideration of fractional operators. However in all the considered cases, it

is possible to exploit the peculiarity of problems in order to avoid fractional derivatives, so that the relevant solutions are always expressed in terms of suitable series expansions for which a general convergence criterion is proved.

The convergence conditions on the data expressed by this criterion are not necessary. It should be interesting to find more general conditions to be satisfied by the data in order to ensure convergence for the formal expansions, but the main motivation of this article is to show the relevance of the H-KdF polynomials in the explicit solutions of the considered problems.

We recall that, in the two-dimensional case, the H-KdF polynomials are closely related to the classical Hermite ones, and consequently to the Gauss' normal distribution which is the weight of the Hermite polynomials.

We conjectured in [5] that this was the hidden motivation of the possibility to express the solutions of all the canonical BVP.

In any case, the importance of the H-KdF polynomials in the framework of the multidimensional or multi-index special functions was recently recognized (see e.g. [7], [4], [10]), and their importance even in the solution of BVP seems to be another property which is worthy to be noted (see e.g. [5], [2], [15]).

2 Hermite-Kampé de Fériet polynomials

We recall the definitions of the H-KdF polynomials, in the two-dimensional case.

Definition 2.1 *The H-KdF polynomials in two variables* $H_m^{(2)}(x,y)$ *are defined by*

$$H_m^{(2)}(x,y) := m! \sum_{n=0}^{\left[\frac{m}{2}\right]} \frac{y^n x^{m-2n}}{n!(m-2n)!} \tag{2.1}$$

Definition 2.2 *The H-KdF polynomials in two variables* $H_m^{(j)}(x,y)$ *are defined by*

$$H_m^{(j)}(x,y) := m! \sum_{n=0}^{\left[\frac{m}{j}\right]} \frac{y^n x^{m-jn}}{n!(m-jn)!} \tag{2.2}$$

In a number of articles by G. Dattoli et al., (see e.g. [7], [8],[9]), by using the so called *monomiality principle*, the following properties for the two-variable H-KdF polynomials $H_m^{(j)}(x,y), \quad j \geq 2$ have been recovered.

- Operational definition

$$H_m^{(j)}(x,y) = e^{y\frac{\partial^j}{\partial x^j}} x^m = \left(x + jy\frac{\partial^{j-1}}{\partial x^{j-1}}\right)^m (1). \tag{2.3}$$

- Generating function

$$\sum_{n=0}^{\infty} H_n^{(j)}(x,y)\ \frac{t^n}{n!} = e^{xt+yt^j}. \tag{2.4}$$

In the case when $j=2$, (see [19]), the H-KdF polynomials $H_m^{(2)}(x,y)$ admit the following

- Integral representation

$$H_m^{(2)}(x,y) = \frac{1}{2\sqrt{\pi y}} \int_{-\infty}^{+\infty} \xi^m e^{-\frac{(x-\xi)^2}{4y}} d\xi, \tag{2.5}$$

which is a particular case of the so called Gauss-Weierstrass (or Poisson) transform [18].

3 Hyperbolic and circular functions of the derivative operator

The two-variable H-KdF polynomials allow us to define, in a constructive way, the hyperbolic and circular functions of the derivative operator.

Definition 3.1 *For any* $j \geq 1$

$$\cosh(yD^j) f(x) := \sum_{m=0}^{\infty} a_m \mathcal{K}_m^{(j)}(x,y), \tag{3.1}$$

where

$$\mathcal{K}_m^{(j)}(x,y) = \mathcal{E}_y\left(H_m^{(j)}(x,y)\right) \tag{3.2}$$

and $\mathcal{E}_y(\cdot)$ *denotes the* even part*, with respect to the* y *variable, of the considered H-KdF polynomial.*

Definition 3.2 *For any* $j \geq 1$

$$\sinh(yD^j) f(x) := \sum_{m=0}^{\infty} a_m \mathcal{S}_m^{(j)}(x,y) \tag{3.3}$$

where

$$\mathcal{S}_m^{(j)}(x,y) = \mathcal{O}_y\left(H_m^{(j)}(x,y)\right) \tag{3.4}$$

and $\mathcal{O}_y(\cdot)$ *denotes the* odd part*, with respect to* y*, of the considered H-KdF polynomial.*

The properties of the hyperbolic and circular functions of the derivative operator, and the relevant generalization to the pseudo-hyperbolic and pseudo-circular functions (see [11], [17], [3]), were considered in [6].

4 Convergence results

In this section we will recall an uniform estimate, with respect to j, for the convergence of series involving the H-KdF polynomials $H_n^{(j)}(x,y)$:

$$\sum_{n=0}^{\infty} a_n \, H_n^{(j)}(x,y). \tag{4.1}$$

Theorem 4.1 *For every* $j \geq 2$, $-\infty < x < +\infty$, $-\infty < y < +\infty$, $n = 0, 1, 2, \ldots$, *the following estimate holds true:*

$$|H_n^{(j)}(x,y)| \leq n! \; exp\left\{|x| + |y|\right\}. \tag{4.2}$$

The proof is derived by the same method used in the book of Widder [19], p. 166, for the case $j = 2$. A deep analysis of the convergence condition for the series (4.1), in the case $j = 2$, is performed in this book, however the relevant estimates can be only partially extended to the general case, since many of them are based on the integral representation (2.5). Unfortunately, if $j > 2$, an integral representation generalizing the Gauss-Weierstrass transform is known only when $j = 2q$, q being an odd number (see [13], [14]).

Considering, as boundary (or initial) data, analytic functions $f(x) = \sum_{n=0}^{\infty} a_n x^n$ whose coefficients a_n tend to zero sufficiently fast, we can prove the following theorem:

Theorem 4.2 *Suppose there exists a number* $\alpha > 1$, *such that the coefficients* a_n *satisfy the following estimate:*

$$|a_n| = O\left(\frac{1}{n^{\alpha} n!}\right), \tag{4.3}$$

then, for every j, *the series expansion* (4.1) *is absolutely and uniformly convergent in every bounded region of the* (x,y)*-plane.*

Proof. The result immediately follows from the estimate

$$\left|\sum_{n=0}^{\infty} a_n \, H_n^{(j)}(x,y)\right| \leq \sum_{n=0}^{\infty} |a_n| \left|H_n^{(j)}(x,y)\right| \leq e^{|x|+|y|} \sum_{n=0}^{\infty} |a_n| n!, \tag{4.4}$$

considering that the last series is convergent by condition (4.3).

Remark 4.1 The condition (4.3) include analytic functions with polynomial growth at infinity, but not the exponential function e^x, whereas, in the case $j = 2$, the book of Widder [19] includes all functions belonging to the so called Huygens class H^0.

Remark 4.2 Obviously the above convergence conditions are sufficient, but not necessary, and it should be interesting to extend it in order to include wider classes of boundary (or initial) data. However, by the point of view of the Applied Analysis, the considered conditions cover many realistic situations, since data are usually expressed by bounded functions, vanishing at infinity, and negligible outside a suitable bounded interval.

In the following we show explicit solutions of the canonical BVP of parabolic, hyperbolic or elliptic type problems, starting from the three-dimensional case, in order to introduce our results in a more friendly way. The relevant extensions to the n-dimensional case are almost straightforward, since the demonstration methods are exactly the same. Consequently, the general results are given without proofs.

5 The two-dimensional heat equation

We start considering the two-dimensional heat equation. Assuming that the initial condition $q(x,y)$ is expressed by an analytic function, and using an operational approach, we find the explicit solution in terms of the H-KdF polynomials. A convergence result is given generalizing the Theorem 4.2.

Theorem 5.1 *Consider the two-dimensional heat equation*

$$\begin{cases} \dfrac{\partial S}{\partial t} = \dfrac{\partial^2 S}{\partial x^2} + \dfrac{\partial^2 S}{\partial y^2} \quad \text{in the half} - \text{space} \quad t > 0, \\ \\ S(x,y,0) = q(x,y) \end{cases} \tag{5.1}$$

with analytic initial condition. Then, the operational solution is given by:

$$S(x,y,t) = e^{t(D_x^2 + D_y^2)} q(x,y), \tag{5.2}$$

where, by definition: $D_x = \dfrac{\partial}{\partial x}$ *and* $D_y = \dfrac{\partial}{\partial y}$.

Proof. We have, indeed:

$$\frac{\partial S}{\partial t} = \left(\frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} \right) e^{t(D_x^2 + D_y^2)} q(x,y) = \frac{\partial^2 S}{\partial x^2} + \frac{\partial^2 S}{\partial y^2}$$

and the initial condition, when $t = 0$, is trivially satisfied.
Considering the Taylor expansion of $q(x,y)$:

$$q(x,y) = \sum_{\substack{n=0 \\ |\alpha| \le n}}^{\infty} a_\alpha \, x^{\alpha_1} \, y^{\alpha_2} \tag{5.3}$$

(where we used the multi-index notation: $\alpha = (\alpha_1, \alpha_2),\ |\alpha| := \alpha_1 + \alpha_2$), it is possible to write down explicitly the solution (5.2).
Consider first the case when $q(x,y) = x^p y^q$. Then:

$$e^{t(D_x^2+D_y^2)}x^p y^q = (x+2tD_x)^p(y+2tD_y)^q(1) = (x+2tD_x)^p H_q^{(2)}(y,t) =$$

$$= \sum_{k=0}^{p}\binom{p}{k}H_{p-k}^{(2)}(x,t)(2tD_x)^k H_q^{(2)}(y,t) = H_p^{(2)}(x,t)H_q^{(2)}(y,t)$$

where we used the Hausdorff and the Burchnall identity (see [7]).
Consequently, for the analytic initial condition (5.3), the solution will be expressed by:

$$S(x,y,t) = e^{t(D_x^2+D_y^2)}\sum_{\substack{n=0\\|\alpha|\le n}}^{\infty} a_\alpha\, x^{\alpha_1}\, y^{\alpha_2} = \tag{5.4}$$

$$= \sum_{\substack{n=0\\|\alpha|\le n}}^{\infty} a_\alpha\,(x+2tD_x)^{\alpha_1}(y+2tD_y)^{\alpha_2}(1) = \sum_{\substack{n=0\\|\alpha|\le n}}^{\infty} a_\alpha\, H_{\alpha_1}^{(2)}(x,t)\, H_{\alpha_2}^{(2)}(y,t).$$

A sufficient condition for the convergence of the expansion (5.4) is expressed by the following result, generalizing Theorem 4.2:

Theorem 5.2 *Suppose that there exist numbers* $\xi_1 > 1,\ \xi_2 > 1$ *such that the coefficients* c_γ *of the series expansion*

$$\sum_{\substack{n=0\\|\gamma|\le n}}^{\infty} c_\gamma H_{\gamma_1}^{(2)}(x,t)H_{\gamma_2}^{(2)}(y,t) \tag{5.5}$$

satisfy

$$|c_\gamma| = O\left(\frac{1}{\gamma_1^{\xi_1}\gamma_1!}\,\frac{1}{\gamma_2^{\xi_2}\gamma_2!}\right), \tag{5.6}$$

then the series (5.5) *is absolutely and uniformly convergent in every bounded region of the* (x,y,t)*-space.*

Proof. Using the estimate (4.2) of Theorem 4.1:

$$\left|H_n^{(j)}(x,t)\right| \le n!\exp\{|x|+|t|\}\ ,$$

we deduce:

$$\left|\sum_{\substack{n=0\\|\gamma|\le n}}^{\infty} c_\gamma H_{\gamma_1}^{(2)}(x,t)H_{\gamma_2}^{(2)}(y,t)\right| \le \sum_{\substack{n=0\\|\gamma|\le n}}^{\infty} |c_\gamma|\left|H_{\gamma_1}^{(2)}(x,t)H_{\gamma_2}^{(2)}(y,t)\right| \le \sum_{\substack{n=0\\|\gamma|\le n}}^{\infty} |c_\gamma|\gamma_1!\,\gamma_2!\,e^{|x|+|y|+2|t|}$$

so that, since by hypothesis the last series is convergent, the series expansion (5.5) satisfies the Weierstrass test in every bounded region of the (x,y,t)-space.

6 The general heat equation

Denote by $\underline{x}$ the $(n-1)$-dimensional vector $\underline{x} \equiv (x_1, x_2, \ldots, x_{n-1})$, and consider the corresponding $(n-1)$-dimensional heat equation, with analytic initial condition $q(\underline{x})$. Putting $\Delta := \partial^2/\partial x_1^2 + \partial^2/\partial x_2^2 + \ldots + \partial^2/\partial x_{n-1}^2$, the result of the preceding section can be easily generalized as follows.

Theorem 6.1 *The operational solution of the problem*

$$\begin{cases} \dfrac{\partial S}{\partial t} = \Delta S, \quad \text{in the half} - \text{space} \quad t > 0, \\ \\ S(\underline{x}, 0) = q(\underline{x}), \end{cases} \tag{6.1}$$

is given by

$$S(\underline{x}, t) = e^{t\Delta} q(\underline{x}). \tag{6.2}$$

If $q(\underline{x}) = x_1^{\alpha_1} \cdots x_{n-1}^{\alpha_{n-1}}$, then the explicit solution of the problem (6.2) is given by

$$\begin{aligned} S(\underline{x}, t) &= e^{t\Delta} x_1^{\alpha_1} \cdots x_{n-1}^{\alpha_{n-1}} = \\ &= (x_1 + 2tD_{x_1})^{\alpha_1} \cdots \left(x_{n-1} + 2tD_{x_{n-1}}\right)^{\alpha_{n-1}} (1) = \\ &= H_{\alpha_1}^{(2)}(x_1, t) \cdots H_{\alpha_{n-1}}^{(2)}(x_{n-1}, t). \end{aligned}$$

For a general analytic initial condition

$$q(\underline{x}) = \sum_{\substack{n=0 \\ |\alpha| \le n}}^{\infty} a_\alpha \, \underline{x}^\alpha,$$

the explicit solution is expressed by the series expansion (6.1):

$$\begin{aligned} S(\underline{x}, t) &= e^{t\Delta} \sum_{\substack{n=0 \\ |\alpha| \le n}}^{\infty} a_\alpha \, \underline{x}^\alpha = \\ &= \sum_{\substack{n=0 \\ |\alpha| \le n}}^{\infty} a_\alpha \, (x_1 + 2tD_{x_1})^{\alpha_1} \cdots \left(x_{n-1} + 2tD_{x_{n-1}}\right)^{\alpha_{n-1}} (1) = \\ &= \sum_{\substack{n=0 \\ |\alpha| \le n}}^{\infty} a_\alpha \, H_{\alpha_1}^{(2)}(x_1, t) \cdots H_{\alpha_{n-1}}^{(2)}(x_{n-1}, t). \end{aligned} \tag{6.3}$$

A convergence condition is given by the theorem

Theorem 6.2 *If there exist* $n-1$ *numbers* $\xi_i > 1$, $(i = 1, 2, \ldots, n-1)$, *such that the coefficients of the series* (6.3) *satisfy the condition:*

$$|a_\alpha| = O\left(\frac{1}{\alpha_1^{\xi_1}\alpha_1!}\cdots\frac{1}{\alpha_{n-1}^{\xi_{n-1}}\alpha_{n-1}!}\right) \tag{6.4}$$

then the series is absolutely and uniformly convergent in every bounded region of the $(\underline{x}, t)$ *space.*

7 The canonical hyperbolic problem in $\mathbf{R^3}$

In this section we find explicit solutions for the canonical hyperbolic problem in $\mathbf{R^3}$, with analytic initial conditions. Convergence results are also derived.

Theorem 7.1 *Consider the hyperbolic problem in* $\mathbf{R^3}$, *with analytic initial data* $q(x, y)$ *and* $v(x, y)$*:*

$$\begin{cases} \dfrac{\partial^2 S}{\partial t^2} = \dfrac{\partial^2 S}{\partial x^2} + \dfrac{\partial^2 S}{\partial y^2} & \text{in the half} - \text{space} \quad t > 0, \\ S(x, y, 0) = q(x, y), \qquad \dfrac{\partial}{\partial t} S(x, y, 0) = v(x, y). \end{cases} \tag{7.1}$$

The operational solution of the problem (7.1) *is given by:*

$$S(x, y, t) = \cosh\left(t\left(\frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2}\right)^{\frac{1}{2}}\right) q(x, y) + \sinh\left(t\left(\frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2}\right)^{\frac{1}{2}}\right) w(x, y) \tag{7.2}$$

where we put, by definition: $w(x, y) := \left(\dfrac{\partial^2}{\partial x^2} + \dfrac{\partial^2}{\partial y^2}\right)^{-\frac{1}{2}} v(x, y)$.

Proof. The initial conditions are trivially satisfied, furthermore, by differentiating two times with respect to t equation (7.2), we find:

$$\begin{aligned} \frac{\partial S}{\partial t} &= \left(\frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2}\right)^{\frac{1}{2}} \sinh\left(t\left(\frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2}\right)^{\frac{1}{2}}\right) q(x, y) + \\ &+ \left(\frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2}\right)^{\frac{1}{2}} \cosh\left(t\left(\frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2}\right)^{\frac{1}{2}}\right) w(x, y) \end{aligned}$$

$$\begin{aligned} \frac{\partial^2 S}{\partial t^2} &= \left(\frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2}\right)\left[\cosh\left(t\left(\frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2}\right)^{\frac{1}{2}}\right) q(x, y) + \sinh\left(t\left(\frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2}\right)^{\frac{1}{2}}\right) w(x, y)\right] \\ &= \frac{\partial^2 S}{\partial x^2} + \frac{\partial^2 S}{\partial y^2}. \end{aligned}$$

By using the Taylor expansion of the initial data:

$$q(x,y) = \sum_{\substack{n=0 \\ |\alpha|\leq n}}^{\infty} a_\alpha \, x^{\alpha_1} \, y^{\alpha_2}, \quad v(x,y) = \sum_{\substack{n=0 \\ |\beta|\leq n}}^{\infty} b_\beta \, x^{\beta_1} \, y^{\beta_2} \, , \tag{7.3}$$

where $\alpha := (\alpha_1, \alpha_2)$, $\beta = (\beta_1, \beta_2)$, and $|\alpha| := \alpha_1 + \alpha_2$, $|\beta| := \beta_1 + \beta_2$, we try to write down explicitly the solution of problem (7.1).

To this aim, we consider first the operational identity:

$$\left(\frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} \right)^m x^p \, y^q = \sum_{k=0}^{m} \binom{m}{k} \left(\frac{\partial}{\partial x} \right)^{2k} \left(\frac{\partial}{\partial y} \right)^{2\,(m-k)} x^p \, y^q \tag{7.4}$$

where we used the commuting property of the operators $\dfrac{\partial^2}{\partial x^2}$ and $\dfrac{\partial^2}{\partial y^2}$.

By using the identities:

$$\begin{aligned} \left(\frac{\partial}{\partial x} \right)^{2k} x^p &= p(p-1)\cdots(p-2k+1)x^{p-2k} = \\ &= \frac{p!}{(p-2k)!} x^{p-2k} \end{aligned}$$

for $p - 2k \geq 0$, or $k \leq \left[\frac{p}{2}\right]$, e

$$\begin{aligned} \left(\frac{\partial}{\partial y} \right)^{2(m-k)} y^q &= q(q-1)\cdots(q-2(m-k)+1)y^{q-2(m-k)} = \\ &= \frac{q!}{(q-2(m-k))!} y^{q-2(m-k)} \end{aligned}$$

for $q - 2(m-k) \geq 0$, or $m \leq \left[\frac{q}{2}\right] + k \leq \left[\frac{q}{2}\right] + \left[\frac{p}{2}\right]$, we find the following equation:

$$\left(\frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} \right)^m x^p \, y^q = \sum_{k=0}^{\left[\frac{p}{2}\right]} \binom{m}{k} \frac{p!}{(p-2k)!} \frac{q!}{(q-2(m-k))!} x^{p-2k} y^{q-2(m-k)} \tag{7.5}$$

where $m \leq \left[\frac{p}{2}\right] + \left[\frac{q}{2}\right]$ (and, by definition, the terms for which $m-k < 0$ or $m-k > \left[\frac{q}{2}\right]$ must be replaced by zero). The above equation (7.5) can be written in the equivalent form:

$$\left(\frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} \right)^m x^p \, y^q = \sum_{\substack{k_1+k_2=m \\ 0\leq k_1\leq\left[\frac{p}{2}\right] \\ 0\leq k_2\leq\left[\frac{q}{2}\right]}} \frac{m!}{k_1!k_2!} \frac{p!}{(p-2k_1)!} \frac{q!}{(q-2k_2)!} x^{p-2k_1} y^{q-2k_2} \tag{7.6}$$

We use now the identity (7.6) in order to represent the solution of the problem (7.1) when the initial data are expressed by the monomials:

$$q(x,y) \,=\, x^p\, y^q, \quad v(x,y) \,=\, x^h\, y^r.$$

In this case, we find:

$$\cosh\left(t\left(\frac{\partial^2}{\partial x^2}+\frac{\partial^2}{\partial y^2}\right)^{\frac{1}{2}}\right)x^p y^q = \sum_{m=0}^{\infty}\frac{t^{2m}}{(2m)!}\left(\frac{\partial^2}{\partial x^2}+\frac{\partial^2}{\partial y^2}\right)^m x^p y^q =$$

$$= \sum_{m=0}^{\left[\frac{p}{2}\right]+\left[\frac{q}{2}\right]}\frac{t^{2m}}{(2m)!}\sum_{\substack{k_1+k_2=m\\ 0\le k_1\le\left[\frac{p}{2}\right]\\ 0\le k_2\le\left[\frac{q}{2}\right]}}\frac{m!}{k_1!k_2!}\frac{p!}{(p-2k_1)!}\frac{q!}{(q-2k_2)!}x^{p-2k_1}y^{q-2k_2} =$$

$$= \sum_{m=0}^{\left[\frac{p}{2}\right]+\left[\frac{q}{2}\right]}\frac{t^m m!}{(2m)!}\left[\sum_{k_1=0}^{\left[\frac{p}{2}\right]}\frac{p!}{k_1!(p-2k_1)!}t^{k_1}x^{p-2k_1}\sum_{k_2=0}^{\left[\frac{q}{2}\right]}\frac{q!}{k_2!(q-2k_2)!}t^{k_2}y^{q-2k_2}\right]_m =$$

$$= \sum_{m=0}^{\left[\frac{p}{2}\right]+\left[\frac{q}{2}\right]}\frac{t^m m!}{(2m)!}\left[H_p^{(2)}(x,t)\cdot H_q^{(2)}(y,t)\right]_m\ , \tag{7.7}$$

where the symbol: $\left[H_p^{(2)}(x,t)\cdot H_q^{(2)}(y,t)\right]_m$ stands for the homogeneous part of degree m, with respect to the t variable, contained into the product: $\left(H_p^{(2)}(x,t)\cdot H_q^{(2)}(y,t)\right)$, i.e. in equation (7.7) we have to erase all the terms for which the exponent of t appearing in the product $\left(H_p^{(2)}(x,t)\cdot H_q^{(2)}(y,t)\right)$ is different from the index m (as a matter of fact, the condition $k_1+k_2=m$ must hold true).

Proceeding in an analogous way we find:

$$\sinh\left(t\left(\frac{\partial^2}{\partial x^2}+\frac{\partial^2}{\partial y^2}\right)^{\frac{1}{2}}\right)x^h\, y^r = \sum_{m=0}^{\infty}\frac{t^{2m+1}}{(2m+1)!}\left(\frac{\partial^2}{\partial x^2}+\frac{\partial^2}{\partial y^2}\right)^{m+\frac{1}{2}}x^h\, y^r$$

$$= \left(\frac{\partial^2}{\partial x^2}+\frac{\partial^2}{\partial y^2}\right)^{\frac{1}{2}}\sum_{m=0}^{\left[\frac{h}{2}\right]+\left[\frac{r}{2}\right]}\frac{t^{2m+1}}{(2m+1)!}\sum_{\substack{k_1+k_2=m\\ 0\le k_1\le\left[\frac{h}{2}\right]\\ 0\le k_2\le\left[\frac{r}{2}\right]}}\frac{m!}{k_1!k_2!}\frac{h!}{(h-2k_1)!}\frac{r!}{(r-2k_2)!}x^{h-2k_1}y^{r-2k_2}$$

$$= \left(\frac{\partial^2}{\partial x^2}+\frac{\partial^2}{\partial y^2}\right)^{\frac{1}{2}}\sum_{m=0}^{\left[\frac{h}{2}\right]+\left[\frac{r}{2}\right]}\frac{t^{m+1}m!}{(2m+1)!}\left[\sum_{k_1=0}^{\left[\frac{h}{2}\right]}\frac{h!}{k_1!(h-2k_1)!}t^{k_1}x^{h-2k_1}\sum_{k_2=0}^{\left[\frac{r}{2}\right]}\frac{r!}{k_2!(r-2k_2)!}t^{k_2}y^{r-2k_2}\right]_m$$

$$= \left(\frac{\partial^2}{\partial x^2}+\frac{\partial^2}{\partial y^2}\right)^{\frac{1}{2}}\sum_{m=0}^{\left[\frac{h}{2}\right]+\left[\frac{r}{2}\right]}\frac{t^{m+1}m!}{(2m+1)!}\left[H_h^{(2)}(x,t)\cdot H_r^{(2)}(y,t)\right]_m\ , \tag{7.8}$$

where the symbol $\left[H_h^{(2)}(x,t)\cdot H_r^{(2)}(y,t)\right]_m$ denotes the homogeneous part of degree m, with respect to the t variable, contained into the product: $\left(H_h^{(2)}(x,t)\cdot H_r^{(2)}(y,t)\right)$.

As a consequence, in this case, the solution of the problem (7.1) becomes:

$$
\begin{aligned}
S(x,y,t) &= \cosh\left(t\left(\frac{\partial^2}{\partial x^2}+\frac{\partial^2}{\partial y^2}\right)^{\frac{1}{2}}\right)x^p y^q \;+\; \sinh\left(t\left(\frac{\partial^2}{\partial x^2}+\frac{\partial^2}{\partial y^2}\right)^{\frac{1}{2}}\right)\left(\frac{\partial^2}{\partial x^2}+\frac{\partial^2}{\partial y^2}\right)^{-\frac{1}{2}} x^h y^r \\
&= \sum_{m=0}^{\left[\frac{p}{2}\right]+\left[\frac{q}{2}\right]} \frac{t^{2m}}{(2m)!} \sum_{\substack{k_1+k_2=m \\ 0\le k_1\le\left[\frac{p}{2}\right] \\ 0\le k_2\le\left[\frac{q}{2}\right]}} \frac{m!}{k_1!k_2!}\frac{p!}{(p-2k_1)!}\frac{q!}{(q-2k_2)!}x^{p-2k_1}y^{q-2k_2} + \\
&+ \sum_{m=0}^{\left[\frac{h}{2}\right]+\left[\frac{r}{2}\right]} \frac{t^{2m+1}}{(2m+1)!} \sum_{\substack{k_1+k_2=m \\ 0\le k_1\le\left[\frac{h}{2}\right] \\ 0\le k_2\le\left[\frac{r}{2}\right]}} \frac{m!}{k_1!k_2!}\frac{h!}{(h-2k_1)!}\frac{r!}{(r-2k_2)!}x^{h-2k_1}y^{r-2k_2} = \\
&= \sum_{m=0}^{\left[\frac{p}{2}\right]+\left[\frac{q}{2}\right]} \frac{t^m m!}{(2m)!}\left[H_p^{(2)}(x,t)\cdot H_q^{(2)}(y,t)\right]_m + \sum_{m=0}^{\left[\frac{h}{2}\right]+\left[\frac{r}{2}\right]} \frac{t^{m+1} m!}{(2m+1)!}\left[H_h^{(2)}(x,t)\cdot H_r^{(2)}(y,t)\right]_m .
\end{aligned}
$$

Consider now the case of general analytic data (7.3). Then the solution of the problem (7.1) becomes:

$$
\begin{aligned}
S(x,y,t) &= \cosh\left(t\left(\frac{\partial^2}{\partial x^2}+\frac{\partial^2}{\partial y^2}\right)^{\frac{1}{2}}\right)\sum_{\substack{n=0 \\ |\alpha|\le n}}^{\infty} a_\alpha\, x^{\alpha_1}\, y^{\alpha_2} + \\
&+ \sinh\left(t\left(\frac{\partial^2}{\partial x^2}+\frac{\partial^2}{\partial y^2}\right)^{\frac{1}{2}}\right)\left(\frac{\partial^2}{\partial x^2}+\frac{\partial^2}{\partial y^2}\right)^{-\frac{1}{2}}\sum_{\substack{n=0 \\ |\beta|\le n}}^{\infty} b_\beta\, x^{\beta_1}\, y^{\beta_2} = \\
&= \sum_{\substack{n=0 \\ |\alpha|\le n}}^{\infty} a_\alpha \cosh\left(t\left(\frac{\partial^2}{\partial x^2}+\frac{\partial^2}{\partial y^2}\right)^{\frac{1}{2}}\right) x^{\alpha_1}\, y^{\alpha_2} + \\
&+ \left(\frac{\partial^2}{\partial x^2}+\frac{\partial^2}{\partial y^2}\right)^{-\frac{1}{2}}\sum_{\substack{n=0 \\ |\beta|\le n}}^{\infty} b_\beta \sinh\left(t\left(\frac{\partial^2}{\partial x^2}+\frac{\partial^2}{\partial y^2}\right)^{\frac{1}{2}}\right) x^{\beta_1}\, y^{\beta_2} =
\end{aligned}
$$

$$
\begin{aligned}
&= \sum_{\substack{n=0\\|\alpha|\le n}}^{\infty} a_\alpha \sum_{m=0}^{\left[\frac{\alpha_1}{2}\right]+\left[\frac{\alpha_2}{2}\right]} \frac{t^{2m}}{(2m)!} \sum_{\substack{k_1+k_2=m\\0\le k_1\le\left[\frac{\alpha_1}{2}\right]\\0\le k_2\le\left[\frac{\alpha_2}{2}\right]}} \frac{m!}{k_1!k_2!}\frac{\alpha_1!}{(\alpha_1-2k_1)!}\frac{\alpha_2!}{(\alpha_2-2k_2)!}x^{\alpha_1-2k_1}y^{\alpha_2-2k_2} + \\
&+ \sum_{\substack{n=0\\|\beta|\le n}}^{\infty} b_\beta \sum_{m=0}^{\left[\frac{\beta_1}{2}\right]+\left[\frac{\beta_2}{2}\right]} \frac{t^{2m+1}}{(2m+1)!} \sum_{\substack{k_1+k_2=m\\0\le k_1\le\left[\frac{\beta_1}{2}\right]\\0\le k_2\le\left[\frac{\beta_2}{2}\right]}} \frac{m!}{k_1!k_2!}\frac{\beta_1!}{(\beta_1-2k_1)!}\frac{\beta_2!}{(\beta_2-2k_2)!}x^{\beta_1-2k_1}y^{\beta_2-2k_2} = \\
&= \sum_{\substack{n=0\\|\alpha|\le n}}^{\infty} a_\alpha \sum_{m=0}^{\left[\frac{\alpha_1}{2}\right]+\left[\frac{\alpha_2}{2}\right]} \frac{t^m m!}{(2m)!}\left[H^{(2)}_{\alpha_1}(x,t)\cdot H^{(2)}_{\alpha_2}(y,t)\right]_m + \\
&+ \sum_{\substack{n=0\\|\beta|\le n}}^{\infty} b_\beta \sum_{m=0}^{\left[\frac{\beta_1}{2}\right]+\left[\frac{\beta_2}{2}\right]} \frac{t^{m+1} m!}{(2m+1)!}\left[H^{(2)}_{\beta_1}(x,t)\cdot H^{(2)}_{\beta_2}(y,t)\right]_m ,
\end{aligned}
\tag{7.9}
$$

where we used the same notations of equations (7.7)-(7.8).

A convergence condition for the series expansions in equation (7.9) is expressed by the theorem:

Theorem 7.2 *Suppose there exist numbers* $\xi_i > 1$ *and* $\rho_i > 1$, $i = 1, 2$ *such that the coefficients* a_α *and* b_β *of the series in equation* (7.9) *satisfy the condition:*

$$
|a_\alpha| = O\left(\frac{1}{\alpha_1^{\xi_1}\alpha_1!}\frac{1}{\alpha_2^{\xi_2}\alpha_2!}\right), \qquad |b_\beta| = O\left(\frac{1}{\beta_1^{\rho_1}\beta_1!}\frac{1}{\beta_2^{\rho_2}\beta_2!}\right) \tag{7.10}
$$

then the series are absolutely and uniformly convergent in every bounded region of the (x, y, t)*-space.*

Proof. We consider the first of the series expansions in equation (7.9), since the result can be proved in the same way for the second one.

By using the estimate:

$$
\left|\sum_{m=0}^{\left[\frac{\alpha_1}{2}\right]+\left[\frac{\alpha_2}{2}\right]} \frac{t^m m!}{(2m)!}\right| \le \sum_{m=0}^{\infty} \frac{|t|^m m!}{(2m)!}
$$

since

$$
\frac{m!}{(2m)!} \le \frac{1}{m!}
$$

we find:

$$\left|\sum_{m=0}^{\left[\frac{\alpha_1}{2}\right]+\left[\frac{\alpha_2}{2}\right]} \frac{t^m m!}{(2m)!}\right| \leq \sum_{m=0}^{\infty} \frac{|t|^m}{(m)!} = \exp\{|t|\}.$$

Furthermore, by exploiting the estimate (4.2) for the H-KdF polynomials, we find:

$$\left|H^{(2)}_{\alpha_1}(|x|,|t|)\cdot H^{(2)}_{\alpha_2}(|y|,|t|)\right| \leq \alpha_1!\alpha_2!\exp\{|x|+2|t|+|y|\}$$

We have so derived the estimate:

$$\sum_{\substack{n=0\\|\alpha|\leq n}}^{\infty} |a_\alpha| \left|\sum_{m=0}^{\left[\frac{\alpha_1}{2}\right]+\left[\frac{\alpha_2}{2}\right]} \frac{t^m m!}{(2m)!}\left[H^{(2)}_{\alpha_1}(x,t)\cdot H^{(2)}_{\alpha_2}(y,t)\right]_m\right| \leq \exp\{|x|+3|t|+|y|\}\sum_{\substack{n=0\\|\alpha|\leq n}}^{\infty} |a_\alpha|\alpha_1!\alpha_2!$$

so that, since by hypothesis the last series are convergent, the series expansion (7.9) satisfies the Weierstrass test in every bounded region of the (x,y,t)-space.

We conclude as follows: *supposing that the coefficients of the analytic initial data of problem* (7.1) *satisfy conditions* (7.10)*, then the solution of the same problem can be expressed by the series expansion* (7.9)*, in terms of the H-KdF polynomials, the series being absolutely and uniformly convergent in every bounded region of the* (x,y,t)*-space.*

8 The hyperbolic problem in $\mathbf{R^n}$

Putting:

$$\underline{x} = (x_1, x_2, \ldots, x_{n-1}),$$

we introduce the multi-index notation:

$$\alpha := (\alpha_1, \ldots, \alpha_{n-1})$$

$$|\alpha| := \sum_{i=1}^{n-1} \alpha_i$$

$$\underline{x}^\alpha := x_1^{\alpha_1}\cdots x_{n-1}^{\alpha_{n-1}},$$

and furthermore we set:

$$\Delta := \frac{\partial^2}{\partial x_1^2} + \frac{\partial^2}{\partial x_2^2} + \ldots + \frac{\partial^2}{\partial x_{n-1}^2}$$

The results of the preceding section can be generalized as follows:

Theorem 8.1 *Consider the hyperbolic problem in* $\mathbf{R^n}$*, with analytic initial data* $q(\underline{x})$ *and* $v(\underline{x})$*:*

$$\begin{cases} \dfrac{\partial^2 S}{\partial t^2} = \Delta\, S \quad \text{in the half} - \text{space} \quad t > 0, \\ S(\underline{x}, 0) = q(\underline{x}), \qquad \dfrac{\partial}{\partial t} S(\underline{x}, 0) = v(\underline{x}). \end{cases} \tag{8.1}$$

The operational solution of the problem (8.1) *is given by:*

$$S(\underline{x}, t) = \cosh\left(t\Delta^{\frac{1}{2}}\right) q(\underline{x}) + \sinh\left(t\Delta^{\frac{1}{2}}\right) w(\underline{x}) \tag{8.2}$$

where, by definition, we put: $w(\underline{x}) := \Delta^{-\frac{1}{2}} v(\underline{x})$.

Considering the analytic initial data, in compact notation:

$$q(\underline{x}) = \sum_{\substack{r=0 \\ |\alpha|\leq r}}^{\infty} a_\alpha\, \underline{x}^\alpha, \qquad v(\underline{x}) = \sum_{\substack{r=0 \\ |\beta|\leq r}}^{\infty} b_\beta\, \underline{x}^\beta \tag{8.3}$$

and generalizing equation (7.6):

$$\Delta^m \underline{x}^\alpha = \sum_{\substack{k_1+\ldots+k_{n-1}=m \\ 0\leq k_i\leq\left[\frac{\alpha_i}{2}\right], i=1:n-1}} \frac{m!}{k_1!\cdots k_{n-1}!} \frac{\alpha_1!}{(\alpha_1-2k_1)!} x_1^{\alpha_1-2k_1} \cdots \frac{\alpha_{n-1}!}{(\alpha_{n-1}-2k_{n-1})!} x_{n-1}^{\alpha_{n-1}-2k_{n-1}} \tag{8.4}$$

where $m \leq \sum_{j=1}^{n-1} \left[\frac{\alpha_j}{2}\right]$, we find the explicit solution in the form:

$$S(\underline{x}, t) = \sum_{\substack{r=0 \\ |\alpha|\leq r}}^{\infty} a_\alpha \cosh\left(t\Delta^{\frac{1}{2}}\right) \underline{x}^\alpha + \Delta^{-\frac{1}{2}} \sum_{\substack{r=0 \\ |\beta|\leq r}}^{\infty} b_\beta \sinh\left(t\Delta^{\frac{1}{2}}\right) \underline{x}^\beta =$$

$$= \sum_{\substack{r=0 \\ |\alpha|\leq r}}^{\infty} a_\alpha \sum_{m=0}^{\sum_{i=1}^{n-1}\left[\frac{\alpha_i}{2}\right]} \frac{t^{2m}}{(2m)!} \sum_{\substack{k_1+\cdots+k_{n-1}=m \\ 0\leq k_i\leq\left[\frac{\alpha_i}{2}\right], i=1:n-1}} \frac{m!}{k_1!\cdots k_{n-1}!} \frac{\alpha_1!}{(\alpha_1-2k_1)!} x_1^{\alpha_1-2k_1} \cdots \frac{\alpha_{n-1}!}{(\alpha_{n-1}-2k_{n-1})!} x_{n-1}^{\alpha_{n-1}-2k_{n-1}} +$$

$$+ \sum_{\substack{r=0 \\ |\beta|\leq r}}^{\infty} b_\beta \sum_{m=0}^{\sum_{i=1}^{n-1}\left[\frac{\beta_i}{2}\right]} \frac{t^{2m+1}}{(2m+1)!} \sum_{\substack{k_1+\cdots+k_{n-1}=m \\ 0\leq k_i\leq\left[\frac{\beta_i}{2}\right], i=1:n-1}} \frac{m!}{k_1!\cdots k_{n-1}!} \frac{\beta_1!}{(\beta_1-2k_1)!} x_1^{\beta_1-2k_1} \cdots \frac{\beta_{n-1}!}{(\beta_{n-1}-2k_{n-1})!} x_{n-1}^{\beta_{n-1}-2k_{n-1}}$$

$$= \sum_{\substack{r=0 \\ |\alpha|\leq r}}^{\infty} a_\alpha \sum_{m=0}^{\sum_{i=1}^{n-1}\left[\frac{\alpha_i}{2}\right]} \frac{t^m m!}{(2m)!} \left[H^{(2)}_{\alpha_1}(x_1,t)\cdots H^{(2)}_{\alpha_{n-1}}(x_{n-1},t)\right]_m +$$

$$+ \sum_{\substack{r=0 \\ |\beta|\leq r}}^{\infty} b_\beta \sum_{m=0}^{\sum_{i=1}^{n-1}\left[\frac{\beta_i}{2}\right]} \frac{t^{m+1} m!}{(2m+1)!} \left[H^{(2)}_{\beta_1}(x_1,t)\cdots H^{(2)}_{\beta_{n-1}}(x_{n-1},t)\right]_m \ , \tag{8.5}$$

where $\left[H^{(2)}_{\alpha_1}(x_1,t)\cdots H^{(2)}_{\alpha_{n-1}}(x_{n-1},t)\right]_m$ denotes the homogeneous part of degree m, with respect to the t variable, contained into the product $\left(H^{(2)}_{\alpha_1}(x_1,t)\cdots H^{(2)}_{\alpha_{n-1}}(x_{n-1},t)\right)$.

The Theorem 4.2 in the n-dimensional case becomes:

Theorem 8.2 *If there exist numbers $\xi_i > 1$, $\rho_i > 1$, $(i = 1, 2, \ldots, n-1)$, such that the coefficients a_α and b_β of the series expansions in equation* (8.5) *satisfy:*

$$|a_\alpha| = O\left(\frac{1}{{\alpha_1}^{\xi_1}\alpha_1!}\cdots\frac{1}{{\alpha_{n-1}}^{\xi_{n-1}}\alpha_{n-1}!}\right), \quad |b_\beta| = O\left(\frac{1}{{\beta_1}^{\rho_1}\beta_1!}\cdots\frac{1}{{\beta_{n-1}}^{\rho_{n-1}}\beta_{n-1}!}\right) \tag{8.6}$$

then the two series are absolutely and uniformly convergent in every bounded region of the $(\underline{x}, t)$-space.

9 The Dirichlet problem for the Laplace equation in $\mathbf{R^3}$

In this section, putting $\Delta := \partial^2/\partial x^2 + \partial^2/\partial y^2 + \partial^2/\partial z^2$, we find explicit solutions for the Dirichlet problem in $\mathbf{R^3}$, with analytic boundary conditions. Convergence results are also derived.

Theorem 9.1 *Consider the Dirichlet problem for the Laplace equation in the half-space $z > 0$, with analytic boundary condition $q(x, y)$, on the real axis $z = 0$:*

$$\begin{cases} \Delta S(x,y,z) = 0, \quad \text{in the half} - \text{space} \quad z > 0, \\ S(x,y,0) = q(x,y) \ . \end{cases} \tag{9.1}$$

The operational solution of the problem (9.1) *is given by:*

$$S(x,y,z) = \cos\left(z\left(\frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2}\right)^{\frac{1}{2}}\right) q(x,y) \tag{9.2}$$

Proof. Note that:

$$\frac{\partial S}{\partial z} = \left(\frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2}\right)^{\frac{1}{2}} \left[-\sin\left(z\left(\frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2}\right)^{\frac{1}{2}}\right) q(x,y)\right]$$

$$\begin{aligned}\frac{\partial^2 S}{\partial z^2} &= \left(\frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2}\right)\left[-\cos\left(z\left(\frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2}\right)^{\frac{1}{2}}\right) q(x,y)\right] \\ &= -\left(\frac{\partial^2 S}{\partial x^2} + \frac{\partial^2 S}{\partial y^2}\right),\end{aligned}$$

and the boundary condition, for $z = 0$, is trivially satisfied.

It is possible again to express the explicit solution of the problem (9.1), when the boundary condition is given by the analytic function:

$$q(x,y) = \sum_{\substack{n=0 \\ |\alpha|\leq n}}^{\infty} a_\alpha\, x^{\alpha_1} y^{\alpha_2}\ , \tag{9.3}$$

where the multi-index notation is used.

Let us consider first the case when

$$q(x,y) = x^p\, y^q.$$

Using equation (7.5) we find the explicit solution in the form:

$$\begin{aligned}&S(x,y,z) = \cos\left(z\left(\frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2}\right)^{\frac{1}{2}}\right) x^p y^q = \\ &= \sum_{m=0}^{\infty} (-1)^m \frac{z^{2m}}{(2m)!}\left[\left(\frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2}\right)^{\frac{1}{2}}\right]^{2m} x^p y^q = \\ &= \sum_{m=0}^{\left[\frac{p}{2}\right]+\left[\frac{q}{2}\right]} (-1)^m \frac{z^{2m}}{(2m)!} \sum_{k=0}^{\left[\frac{p}{2}\right]} \binom{m}{k} \frac{p!}{(p-2k)!} \frac{q!}{(q-2(m-k))!} x^{p-2k} y^{q-2(m-k)} = \\ &= \sum_{m=0}^{\left[\frac{p}{2}\right]+\left[\frac{q}{2}\right]} (-1)^m \frac{z^m m!}{(2m)!} \left[H_p^{(2)}(x,z)\cdot H_q^{(2)}(y,z)\right]_m\ ,\end{aligned} \tag{9.4}$$

where the symbol $\left[H_p^{(2)}(x,z)\cdot H_q^{(2)}(y,z)\right]_m$ denotes the homogeneous part of degree m, with respect to the z variable, contained in the product $\left(H_p^{(2)}(x,z)\cdot H_q^{(2)}(y,z)\right)$,

i.e. all the terms whose index m is different from the exponent of z must be replaced by zero.
Note that, in order to obtain the last equation, we decomposed first z^{2m} in the form $z^{2m} = z^m \cdot z^{k_1} \cdot z^{k_2}\,(k_1 + k_2 = m)$, so that, by using (7.6) which is equivalent to (7.5), it follows:

$$\sum_{m=0}^{\left[\frac{p}{2}\right]+\left[\frac{q}{2}\right]} (-1)^m \frac{z^m}{(2m)!} \sum_{\substack{k_1+k_2=m \\ 0\le k_1\le\left[\frac{p}{2}\right] \\ 0\le k_2\le\left[\frac{q}{2}\right]}} \frac{m!}{k_1!k_2!} \frac{p!}{(p-2k_1)!} z^{k_1} x^{p-2k_1} \frac{q!}{(q-2k_2)!} z^{k_2} y^{q-2k_2}.$$

Using the definition of the H-KdF polynomials:

$$H_p^{(2)}(x,z) = p! \sum_{k_1=0}^{\left[\frac{p}{2}\right]} \frac{z^{k_1} x^{p-2k_1}}{k_1!(p-2k_1)!} \qquad H_q^{(2)}(y,z) = q! \sum_{k_2=0}^{\left[\frac{q}{2}\right]} \frac{z^{k_2} y^{q-2k_2}}{k_2!(q-2k_2)!}$$

we find:

$$\sum_{m=0}^{\left[\frac{p}{2}\right]+\left[\frac{q}{2}\right]} (-1)^m \frac{z^m m!}{(2m)!} \left[p! \sum_{k_1=0}^{\left[\frac{p}{2}\right]} \frac{z^{k_1} x^{p-2k_1}}{k_1!(p-2k_1)!}\, q! \sum_{k_2=0}^{\left[\frac{q}{2}\right]} \frac{z^{k_2} y^{q-2k_2}}{k_2!(q-2k_2)!} \right]_m$$

so that equation (9.4) holds true.
In the general case, when the boundary condition is expressed in the form (9.3), the explicit solution becomes:

$$S(x,y,z) = \cos\left(z \left(\frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} \right)^{\frac{1}{2}} \right) \sum_{\substack{n=0 \\ |\alpha|\le n}}^{\infty} a_\alpha\, x^{\alpha_1}\, y^{\alpha_2} =$$

$$= \sum_{\substack{n=0 \\ |\alpha|\le n}}^{\infty} a_\alpha \cos\left(z \left(\frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} \right)^{\frac{1}{2}} \right) x^{\alpha_1}\, y^{\alpha_2} =$$

$$= \sum_{\substack{n=0 \\ |\alpha|\le n}}^{\infty} a_\alpha \sum_{m=0}^{\left[\frac{\alpha_1}{2}\right]+\left[\frac{\alpha_2}{2}\right]} (-1)^m \frac{z^{2m}}{(2m)} \sum_{k=0}^{\left[\frac{\alpha_1}{2}\right]} \binom{m}{k} \frac{\alpha_1!}{(\alpha_1-2k)!} \frac{\alpha_2!}{(\alpha_2-2(m-k))!} x^{\alpha_1-2k} y^{\alpha_2-2(m-k)} =$$

$$= \sum_{\substack{n=0 \\ |\alpha|\le n}}^{\infty} a_\alpha \sum_{m=0}^{\left[\frac{\alpha_1}{2}\right]+\left[\frac{\alpha_2}{2}\right]} (-1)^m \frac{z^m m!}{(2m)!} \left[H_{\alpha_1}^{(2)}(x,z) \cdot H_{\alpha_2}^{(2)}(y,z) \right]_m . \tag{9.5}$$

A convergence condition, for the series expansion in equation (9.5), is given by the theorem:

Theorem 9.2 *If there exist numbers* $\xi_1 > 1$, $\xi_2 > 1$ *such that the coefficients* a_α *of the series* (9.5) *satisfy the estimate:*

$$|a_\alpha| = O\left(\frac{1}{\alpha_1^{\xi_1}\alpha_1!}\,\frac{1}{\alpha_2^{\xi_2}\alpha_2!}\right), \tag{9.6}$$

then the series (9.5) *is absolutely and uniformly convergent in every bounded region of the* (x, y, z)*-space.*

Proof.
The result follows from the estimates:

$$\left|\sum_{m=0}^{\left[\frac{\alpha_1}{2}\right]+\left[\frac{\alpha_2}{2}\right]} (-1)^m \frac{z^m m!}{(2m)!}\right| \leq \sum_{m=0}^{\infty} \frac{|z|^m}{(m)!} = \exp\{|z|\}.$$

$$\left|H^{(2)}_{\alpha_1}(|x|, |z|) \cdot H^{(2)}_{\alpha_2}(|y|, |z|)\right| \leq \alpha_1!\alpha_2! \exp\{|x| + 2|z| + |y|\},$$

hence the series (9.5) is dominated by

$$\exp\{|x| + 3|z| + |y|\} \sum_{\substack{n=0 \\ |\alpha|\leq n}}^{\infty} |a_\alpha|\alpha_1!\alpha_2!$$

which is convergent under the considered hypotheses, so that the series (9.5) satisfies the Weierstrass test in every bounded region of the (x, y, z)-space.

We conclude: *supposing that the coefficients of the analytic boundary condition* (9.3) *of the Dirichlet problem* (9.1) *satisfy* (9.6)*, then the solution of the same problem can be expressed by the series expansion* (9.5)*, in terms of the H-KdF polynomials, the series being absolutely and uniformly convergent in every bounded region of the* (x, y, z)*- space.*

10 The Dirichlet problem for the Laplace equation in $\mathbf{R^n}$

Consider now the n-dimensional Laplace operator $\Delta := \partial^2/\partial x_1^2 + \partial^2/\partial x_2^2 + \ldots + \partial^2/\partial x_n^2$, in the half-space $x_n > 0$, and the Dirichlet problem with boundary condition on the real x_n axis, expressed by the analytic function $q(\underline{x})$, where we put again $\underline{x} := (x_1, \ldots, x_{n-1})$.
Then:

Theorem 10.1 *The operational solution of the problem*

$$\begin{cases} \Delta_n S(\underline{x}, x_n) = 0, \quad \text{in the half} - \text{space} \quad x_n > 0, \\ S(\underline{x}, 0) = q(\underline{x}). \end{cases} \tag{10.1}$$

is given by

$$S(\underline{x}, x_n) = \cos\left(x_n \Delta^{\frac{1}{2}}\right) q(\underline{x}) \tag{10.2}$$

If the boundary condition is expressed by the function:

$$q(\underline{x}) = \sum_{\substack{r=0 \\ |\alpha|\leq r}}^{\infty} a_\alpha \, \underline{x}^\alpha$$

(where we used again the compact multi-index notation), then the explicit solution in terms of the H-KdF polynomials is:

$$S(\underline{x}, x_n) = \cos\left(x_n \Delta^{\frac{1}{2}}\right) \sum_{\substack{r=0 \\ |\alpha|\leq r}}^{\infty} a_\alpha \, \underline{x}^\alpha = \sum_{\substack{r=0 \\ |\alpha|\leq r}}^{\infty} a_\alpha \sum_{m=0}^{\infty} (-1)^m \frac{x_n^{2m}}{(2m)!} \Delta^m \underline{x}^\alpha =$$

$$= \sum_{\substack{r=0 \\ |\alpha|\leq r}}^{\infty} a_\alpha \sum_{m=0}^{\sum_{i=1}^{n-1}\left[\frac{\alpha_i}{2}\right]} (-1)^m \frac{x_n^{2m}}{(2m)!} \sum_{\substack{k_1+\cdots+k_{n-1}=m \\ 0\leq k_i \leq \left[\frac{\alpha_i}{2}\right]}} \frac{m!}{k_1! \cdots k_{n-1}!} \frac{\alpha_1!}{(\alpha_1 - 2k_1)!} x_1^{\alpha_1 - 2k_1} \cdots \frac{\alpha_{n-1}!}{(\alpha_{n-1} - 2k_{n-1})!} x_{n-1}^{\alpha_{n-1} - 2k_{n-1}}$$

$$= \sum_{\substack{r=0 \\ |\alpha|\leq r}}^{\infty} a_\alpha \sum_{m=0}^{\sum_{i=1}^{n-1}\left[\frac{\alpha_i}{2}\right]} (-1)^m \frac{x_n^m}{(2m)!} \left[H^{(2)}_{\alpha_1}(x_1, x_n) \cdots H^{(2)}_{\alpha_{n-1}}(x_{n-1}, x_n)\right]_m . \tag{10.3}$$

A convergence condition, for the series expansions in equation (10.3), is given by the theorem:

Theorem 10.2 *If there exist* $n-1$ *numbers* $\xi_i > 1$, $(i = 1, 2, \ldots, n-1)$, *such that the coefficients* a_α *of the series* (10.3) *satisfy the estimate:*

$$|a_\alpha| = O\left(\frac{1}{{\alpha_1}^{\xi_1} \alpha_1!} \cdots \frac{1}{{\alpha_{n-1}}^{\xi_{n-1}} \alpha_{n-1}!}\right), \tag{10.4}$$

then the series (10.3) *are absolutely and uniformly convergent in every bounded region of the* $(\underline{x}, x_n)$*-space.*

Acknowledgements
This article was concluded in the framework of the Italian National Group for Scientific Computation (G.N.C.S.), whose partial support we wish to acknowledge.

The Authors express their sincere gratitude to Dr. G. Dattoli, and Prof. Y. Ben Cheikh for useful discussions and comments.

References

[1] P. Appell, J. Kampé de Fériet, *Fonctions Hypergéométriques et Hypersphériques. Polynômes d'Hermite*, Gauthier-Villars, Paris, 1926.

[2] Y. Ben Cheikh, G. Dattoli, C. Cesarano, On some partial differential equations, *Ann. Univ. Ferrara*, (submitted).

[3] Y. Ben Cheikh, Decomposition of some complex functions with respect to the cyclic group of order n, *Appl. Math. Inform.*, **4**, 30–53 (1999).

[4] G. Bretti, P.E. Ricci, Multidimensional extensions of the Bernoulli and Appell Polynomials, *Taiwanese J. Math.*, **8** (2004), 415–428.

[5] C. Cassisa, P.E. Ricci, I. Tavkhelidze, An operatorial approach to solutions of BVP in the half-plane, *J. Concr. & Appl. Math.*, **1**, 37–62 (2003).

[6] C. Cassisa, P.E. Ricci, I. Tavkhelidze, Operational identities for circular and hyperbolic functions and their generalizations, *Georgian Math. J.*, **10, 45–56** (2003).

[7] G. Dattoli, A. Torre, P.L. Ottaviani, L. Vázquez, Evolution operator equations: integration with algebraic and finite difference methods. Application to physical problems in classical and quantum mechanics, *Riv. Nuovo Cimento*, **2**, 1–133 (1997).

[8] G. Dattoli, A. Torre, C. Carpanese, Operational rules and arbitrary order Hermite generating functions, *J. Math. Anal. Appl.*, **227**, 98–111 (1998).

[9] G. Dattoli, A. Torre, Operational methods and two variable Laguerre polynomials, *Atti Acc. Sc. Torino*, 132, 1–7 (1998).

[10] G. Dattoli, P.E. Ricci, H.M. Srivastava, Two-index multidimensional Gegenbauer polynomials and integral representations, *Math. & Comput. Modelling*, **37**, 283–291 (2003).

[11] A. Erdély, W. Magnus, F. Oberhettinger, F.G. Tricomi, *Higher Trascendental Functions*, McGraw-Hill, New York, 1953.

[12] H.W. Gould, A.T. Hopper, Operational formulas connected with two generalizations of Hermite Polynomials, *Duke Math. J.*, **29**, 51–62 (1962).

[13] D.T. Haimo, C. Market, A reprentation theory for solutions of a higher order heat equation, I, *J. Math. Anal. Appl.*, **168**, 89–107 (1992).

[14] D.T. Haimo, C. Market, A reprentation theory for solutions of a higher order heat equation, II, *J. Math. Anal. Appl.*, **168**, 289–305 (1992).

[15] G. Maroscia, P.E. Ricci, Explicit solutions of multidimensional pseudo-classical BVP in the half-space, *Math. Comput. Modelling*, **40** (2004), 667–698.

[16] H.M. Srivastava, H.L. Manocha, *A treatise on generating functions*, Wiley, New York, 1984.

[17] P.E. Ricci, Le Funzioni Pseudo-Iperboliche e Pseudo-trigonometriche, *Pubbl. Ist. Mat. Appl. Fac. Ingegneria Univ. Stud. Roma*, **12**, 37–49, (1978).

[18] D.V. Widder, *An Introduction to Transform Theory*, Academic Press, New York, 1971.

[19] D.V. Widder, *The Heat Equation*, Academic Press, New York, 1975.

JOURNAL OF CONCRETE AND APPLICABLE MATHEMATICS,VOL.3,NO.1,31-40,2005,COPYRIGHT 2005 EUDOXUS PRESS,LLC

GENERALIZED m-ACCRETIVE MAPPINGS AND VARIATIONAL INCLUSIONS IN BANACH SPACES

NAN-JING HUANG
DEPARTMENT OF MATHEMATICS
SICHUAN UNIVERSITY
CHENGDU, SICHUAN 610064, P. R. CHINA
E-MAIL: NANJINGHUNAG@HOTMAIL.COM

YA-PING FANG
DEPARTMENT OF MATHEMATICS
SICHUAN UNIVERSITY
CHENGDU, SICHUAN 610064, P. R. CHINA
E-MAIL: FABHCN@YAHOO.COM.CN

YEOL JE CHO
DEPARTMENT OF MATHEMATICS EDUCATION AND THE RINS
COLLEGE OF EDUCATION
GYSONGSANG NATIONAL UNIVERSITY
CHINJU 660-701, KOREA
E-MAIL: YJCHO@GSNU.AC.KR

Abstract. Some properties of generalized m-accretive mappings in Banach spaces are proved in this paper. We also introduce and study a new class of variational inclusions with generalized m-accretive mappings and construct a new algorithm for approximating the solution of the variational inclusions with generalized m-accretive mappings in q-uniformly smooth Banach spaces by using the resolvent operator technique.

2000 Mathematics Subject Classification. Primary 47H06; Secondary 49H40.

Key words and phrases. Generalized m-accretive mapping, variational inclusion, resolvent operator, iterative algorithm, convergence.

1. Introduction

Variational inequality theory, as a very effective and powerful tool of the current mathematical technology, has been widely applied to mechanics, physics, optimization and control, economic and transportation equilibrium, and engineering sciences, etc. (see [1]-[6], [8]-[13], [15] and the references therein). Because of its wide applications, the classical variational inequality has been generalized in various directions for the past years. Variational inclusion is an important generalization of variational inequality and has been studied by many authors. We also know that one of the most important and interesting problems in the theory of variational inequality is the development of an efficient and implementable algorithm for solving various variational inequalities and variational inclusions. In recent years, many numerical methods have been developed for solving various classes of variational

inequalities and variational inclusions in Euclidean Spaces or Hilbert spaces, such as the projection method and its variant forms, linear approximation, descent, and Newton's methods. However, few iterative algorithms have been developed for solving variational inequality and inclusion problems in Banach spaces.

Recently, Huang and Fang [7] introduced the concept of a generalized m-accretive mapping, which is a generalization of an m-accretive mapping, and gave the definition of the resolvent operator for a generalized m-accretive mapping in a Banach space.

Motivated and inspired by the results in [2], [7] and [9], in this paper, we prove the Lipschitz continuity of the resolvent operator for a generalized m-accretive mapping in Banach spaces and construct a new iterative algorithm for solving a new class of variational inclusions with generalized m-accretive mappings in q-uniformly smooth Banach spaces by using the resolvent operator technique.

2. Preliminaries

Let X be a real Banach space with the dual space X^* and $\langle \cdot, \cdot \rangle$ be the dual pair between X and X^*. The *generalized duality mapping* $J_q : X \to 2^{X^*}$ is defined by

$$J_q(x) = \{f^* \in X^* : \langle x, f^* \rangle = \|x\|^q \quad \text{and} \quad \|f^*\| = \|x\|^{q-1}\}, \quad x \in X,$$

where $q > 1$ is a constant. In particular, J_2 is the usual normalized duality mapping. It is known that, in general, $J_q(x) = \|x\|^{q-2} J_2(x)$ for all $x \in X$ $(x \neq 0)$ and J_q is single-valued if X^* is strictly convex (see, for example, [14]). If $X = H$ is a Hilbert space, then J_2 becomes the identity mapping of H. In the sequel, we shall denote the single-valued generalized duality mapping by j_q.

The *modulus of smoothness* of X is the function $\rho_X : [0, \infty) \to [0, \infty)$ defined by

$$\rho_X(t) = \sup \left\{ \frac{1}{2}(\|x + y\| + \|x - y\|) - 1 : \|x\| \leq 1, \|y\| \leq t \right\}.$$

A Banach space X is said to be *uniformly smooth* if

$$\lim_{t \to 0} \frac{\rho_X(t)}{t} = 0.$$

The Banach space X is said to be *q-uniformly smooth* if there exists a constant $c > 0$ such that

$$\rho_X \leq ct^q, \quad q > 1.$$

All Hilbert spaces, L_p(or l_p) spaces, $1 < p < \infty$, and the Sobolev spaces $W^{m,p}$, $1 < p < \infty$, are all q-uniformly smooth.

In the study of the characteristic inequalities in q-uniformly smooth Banach spaces, Xu [14] proved the following theorem:

Theorem X. *Let X be a real uniformly smooth Banach space. Then X is q-uniformly smooth if and only if there exists a constant $c_q > 0$ such that, for all $x, y \in X$,*

$$\|x + y\|^q \leq \|x\|^q + q\langle y, j_q(x) \rangle + c_q \|y\|^q.$$

Definition 2.1. A multivalued mapping $A : X \to 2^X$ is said to be
(1) *accretive* if, for any $u, v \in X$, there exists $j_2(u - v) \in J_2(u - v)$ such that

$$\langle x - y, j_2(u - v) \rangle \geq 0, \quad x \in Au, \, y \in Av,$$

(2) *strictly accretive* if, for any $u, v \in X$, there exists $j_2(u-v) \in J_2(u-v)$ such that

$$\langle x-y, j_2(u-v)\rangle \geq 0, \quad x \in Au,\ y \in Av,$$

and the equality holds if and only if $u = v$,

(3) *strongly accretive* if, for any $u, v \in X$, there exists $j_2(u-v) \in J_2(u-v)$ such that

$$\langle x-y, j_2(u-v)\rangle \geq \gamma\|u-v\|^2, \quad x \in Au,\ y \in Av,$$

where $\gamma > 0$ is a constant,

(4) *m-accretive* if A is accretive and $(I+\lambda A)(X) = X$ for all (equivalently, for some) $\lambda > 0$, where I denotes the identity mapping.

Remark 2.1. (1) In Definition 2.1, we can replace j_2 by j_q for all $q > 1$ and the strongly accretive constant γ is independent of q.

(2) If $X = X^* = H$ is a Hilbert space, then (1)–(4) of Definition 2.1 reduce to the definitions of monotonicity, strict monotonicity, strong monotonicity and maximal monotonicity, respectively.

Definition 2.2. A mapping $T : X \to X$ is said to be *Lipschitz continuous* if there exists a constant $k > 0$ such that

$$\|Tu - Tv\| \leq k\|u-v\|, \quad u, v \in X.$$

3. Generalized m-Accretive Mappings

Definition 3.1. A mapping $\eta : X \times X \to X^*$ is said to be

(1) *monotone* if

$$\langle u-v, \eta(u,v)\rangle \geq 0, \quad u, v \in X,$$

(2) *strictly monotone* if

$$\langle u-v, \eta(u,v)\rangle \geq 0, \quad u, v \in X,$$

and the equality holds if and only if $u = v$,

(3) *strongly monotone* if there exists a constant $\delta > 0$ such that

$$\langle u-v, \eta(u,v)\rangle \geq \delta\|u-v\|^2, \quad u, v \in X,$$

(4) *Lipschitz continuous* if there exists a constant $\tau > 0$ such that

$$\|\eta(u,v)\| \leq \tau\|u-v\|, \quad u, v \in X.$$

Note that the strong monotonicity of η implies the strict monotonicity of η.

Definition 3.2. ([7]) Let $\eta : X \times X \to X^*$ be a single-valued mapping and $M : X \to 2^X$ be a multivalued mapping. Then M is said to be

(1) *η-accretive* if

$$\langle x-y, \eta(u,v)\rangle \geq 0, \quad u, v \in X,\ x \in Mu,\ y \in Mv,$$

(2) *strictly η-accretive* if

$$\langle x-y, \eta(u,v)\rangle \geq 0, \quad u, v \in X,\ x \in Mu,\ y \in Mv,$$

and the equality holds if and only if $u = v$,

(3) *strongly η-accretive* if there exists a constant $r > 0$ such that

$$\langle x-y, \eta(u,v)\rangle \geq r\|u-v\|^2, \quad u, v \in X,\ x \in Mu,\ y \in Mv,$$

(4) *generalized m-accretive* if M is η-accretive and $(I+\lambda M)(X)=X$ for all (equivalently, for some) $\lambda>0$.

Remark 3.1. (1) In [7], Huang and Fang introduced the concept of generalized m-accretive mappings in Banach spaces and gave an example of the generalized m-accretive mapping and some properties of the generalized m-accretive mapping.

(2) If $X=X^*=H$ is a Hilbert space, then (1)–(4) of Definition 3.2 reduce to the definitions of η-monotonicity, strict η-monotonicity, strong η-monotonicity and maximal η-monotonicity, respectively.

(3) If X is uniformly smooth and $\eta(x,y)=J_2(x-y)$, then (1)–(4) of Definition 3.2 reduce to the definitions of accretivity, strict accretivity, strong accretivity, and m-accretivity in uniformly smooth Banach spaces, respectively.

Now we prove some properties of the generalized m-accretive mapping in Banach spaces.

Theorem 3.1. *Let $\eta: X\times X\to X^*$ be strictly monotone and $M: X\to 2^X$ be a generalized m-accretive mapping. Then the following conclusions hold:*

(1) $\langle x-y,\eta(u,v)\rangle\geq 0$ *for all* $(y,v)\in Graph(M)$ *implies that* $(x,u)\in Graph(M)$, *where* $Graph(M)=\{(x,u)\in X\times X: x\in Mu\}$.

(2) *The inverse mapping* $(I+\lambda M)^{-1}$ *is single-valued for all* $\lambda>0$.

Proof. Suppose that (1) is false. Then there exists $(x_0,u_0)\notin Graph(M)$ such that

$$\langle x_0-y,\eta(u_0,v)\rangle\geq 0,\quad (y,v)\in Graph(M). \tag{3.1}$$

Since M is generalized m-accretive, we have $(I+\lambda M)(X)=X$. Then there exists $(x_1,u_1)\in Graph(M)$ such that

$$u_1+\lambda x_1=u_0+\lambda x_0. \tag{3.2}$$

It follows from (3.1) and (3.2) that

$$\lambda\langle x_0-x_1,\eta(u_0,u_1)\rangle=\langle u_1-u_0,\eta(u_0,u_1)\rangle\geq 0.$$

This implies that

$$\langle u_0-u_1,\eta(u_0,u_1)\rangle\leq 0.$$

Since η is strictly monotone, we have $u_0=u_1$ and hence, from (3.2), we have $x_1=x_0$. This contradicts the fact $(x_0,u_0)\notin Graph(M)$. Thus (1) is true.

Now we prove (2). For any given $z\in X$ and a constant $\lambda>0$, let $u,v\in(I+\lambda M)^{-1}(z)$. Then $\lambda^{-1}(z-u)\in M(u)$ and $\lambda^{-1}(z-v)\in M(v)$. By η-accretivity of M, we obtain

$$\begin{aligned}0&=\langle z-z,\eta(u,v)\rangle\\&=\lambda\langle\frac{1}{\lambda}(z-u)-\frac{1}{\lambda}(z-v),\eta(u,v)\rangle+\langle u-v,\eta(u,v)\rangle\\&\geq\langle u-v,\eta(u,v)\rangle.\end{aligned}$$

Since η is strictly monotone, we have $u=v$. Thus $(I+\lambda M)^{-1}$ is single-valued. This completes the proof.

Based on Theorem 3.1, we can define the resolvent operator for a generalized m-accretive mapping M as follows:

$$J_\rho^M(z) = (I + \rho M)^{-1}(z), \quad z \in X,$$

where $\rho > 0$ is a constant and $\eta : X \times X \to X^*$ is a strictly monotone mapping.

Theorem 3.2. *Let $\eta : X \times X \to X^*$ be strongly monotone and Lipschitz continuous with constants $\delta > 0$ and $\tau > 0$, respectively, and $M : X \to 2^X$ be a generalized m-accretive mapping. Then the resolvent operator J_ρ^M for M is Lipschitz continuous with constant τ/δ, i.e.,*

$$\|J_\rho^M(u) - J_\rho^M(v)\| \leq \frac{\tau}{\delta}\|u - v\|, \quad u, v \in X. \tag{3.3}$$

Proof. Let u, v be any given points in X. From the definition of J_ρ^M, we have

$$J_\rho^M(u) = (I + \rho M)^{-1}(u), \quad J_\rho^M(v) = (I + \rho M)^{-1}(v),$$

which imply that

$$\frac{1}{\rho}(u - J_\rho^M(u)) \in M(J_\rho^M(u)), \quad \frac{1}{\rho}(v - J_\rho^M(v)) \in M(J_\rho^M(v)).$$

Since M is η-accretive, we obtain

$$\begin{aligned} &\frac{1}{\rho}\langle u - J_\rho^M(u) - (v - J_\rho^M(v)), \eta(J_\rho^M(u), J_\rho^M(v))\rangle \\ &= \frac{1}{\rho}\langle u - v - (J_\rho^M(u) - J_\rho^M(v)), \eta(J_\rho^M(u), J_\rho^M(v))\rangle \\ &\geq 0. \end{aligned}$$

From the above inequality, we have

$$\begin{aligned} \delta\|J_\rho^M(u) - J_\rho^M(v)\|^2 &\leq \langle J_\rho^M(u) - J_\rho^M(v), \eta(J_\rho^M(u), J_\rho^M(v))\rangle \\ &\leq \langle u - v, \eta(J_\rho^M(u), J_\rho^M(v))\rangle \\ &\leq \tau\|u - v\| \cdot \|J_\rho^M(u) - J_\rho^M(v)\|, \end{aligned}$$

which implies that

$$\|J_\rho^M(u) - J_\rho^M(v)\| \leq \frac{\tau}{\delta}\|u - v\|, \quad u, v \in X.$$

This completes the proof.

4. Variational Inclusions

In this section, we shall study a new class of variational inclusions with generalized m-accretive mappings in Banach spaces. We also construct a new iterative algorithm for approximating the solution of this class of variational inclusions with generalized m-accretive mappings in q-uniformly smooth Banach spaces by using the resolvent operator technique.

Let $\eta : X \times X \to X^*$ and $A : X \to X$ be two single-valued mappings and $M : X \to 2^X$ be a generalized m-accretive multivalued mapping. Now we consider the following problem:

Find $u \in X$ such that

$$0 \in A(u) + M(u). \tag{4.1}$$

Remark 4.1. (1) If $X = X^* = H$ is a Hilbert space and $\eta(u, v) = u - v$, then the problem (4.1) becomes the usual variational inclusion with a maximal monotone mapping M.

(2) If $X = X^* = H$ is a Hilbert space, $\eta(u, v) = u - v$ and $M = \partial\varphi$, where $\partial\varphi$ denotes the subdifferential of a proper convex lower semicontinuous function φ on H, then the problem (4.1) becomes the following classical variational inequality:

Find $u \in H$ such that

$$\langle Au, v - u\rangle + \varphi(v) - \varphi(u) \geq 0, \quad v \in H.$$

Lemma 4.1. *For given $u \in X$, u is a solution of the problem (4.1) if and only if*

$$u = J_\rho^M(u - \rho Au),$$

where $\rho > 0$ is a constant.

Proof. The proof directly follows from the definition of J_ρ^M.

Lemma 4.2. *Let $\{a_n\}$, $\{b_n\}$ and $\{c_n\}$ be three sequences of nonnegative real numbers satisfying the following conditions:*

There exists a positive integer n_0 such that

$$a_{n+1} \leq (1 - t_n)a_n + b_n t_n + c_n, \quad n \geq n_0, \tag{4.2}$$

where

$$t_n \in [0, 1], \quad \sum_{n=0}^{\infty} t_n = +\infty, \quad \lim_{n\to\infty} b_n = 0, \quad \sum_{n=0}^{\infty} c_n < +\infty.$$

Then $a_n \to 0$ as $n \to +\infty$.

Proof. Let $\sigma = \inf\{a_n : n \geq n_0\}$. Then $\sigma \geq 0$. Suppose that $\sigma > 0$ and so $a_n \geq \sigma > 0$ for all $n \geq n_0$. It follows from (4.2) that

$$\begin{aligned} a_{n+1} &\leq a_n - \sigma t_n + t_n b_n + c_n \\ &= a_n - (\frac{1}{2}\sigma - b_n)t_n - \frac{1}{2}\sigma t_n + c_n \quad n \geq n_0. \end{aligned} \tag{4.3}$$

Since $b_n \to 0$ as $n \to \infty$, there exists $n_1 \geq n_0$ such that

$$\frac{1}{2}\sigma \geq b_n, \quad n \geq n_1. \tag{4.4}$$

Combining (4.3) and (4.4), then we have

$$a_{n+1} \leq a_n - \frac{1}{2}\sigma t_n + c_n, \quad n \geq n_1,$$

which implies that

$$\frac{1}{2}\sigma \sum_{n=n_1}^{\infty} t_n \leq a_{n_1} + \sum_{n=n_1}^{\infty} c_n < +\infty.$$

This is a contradiction. Therefore, $\sigma = 0$ and so there exists a subsequence $\{a_{n_j}\}$ of $\{a_n\}$ such that $a_{n_j} \to 0$ as $j \to \infty$. It follows from (4.2) that

$$a_{n_j+1} \leq a_{n_j} + b_{n_j} t_{n_j} + c_{n_j}$$

and so $a_{n_j+1} \to 0$ as $j \to \infty$. A simple induction leads to $a_{n_j+k} \to 0$ as $j \to \infty$ for all $k \geq 1$ and this means that $a_n \to 0$ as $n \to \infty$. This completes the proof.

Theorem 4.1. *Let X be a q-uniformly smooth Banach space. Let $\eta : X \times X \to X^*$ be a strongly monotone and Lipschitz continuous mapping with constants δ and τ, respectively, $A : X \to X$ be a strong accretive and Lipschitz continuous single-valued mapping with constants r and s, respectively, and $M : X \to 2^X$ be a generalized m-accretive multivalued mapping. Assume that there exists $\rho > 0$ such that*

$$\tau(1 - rq\rho + c_q s^q \rho^q)^{\frac{1}{q}} < \delta, \tag{4.5}$$

where $c_q > 0$ is the same as in Theorem X. Then the problem (4.1) has a unique solution $u^* \in X$.

Proof. Define a mapping $F : X \to X$ by

$$F(u) = J_\rho^M(u - \rho Au), \quad u \in X. \tag{4.6}$$

It follows from (3.3) and (4.6) that

$$\begin{aligned} \|F(u) - F(v)\| &= \|J_\rho^M(u - \rho Au) - J_\rho^M(v - \rho Av)\| \\ &\leq \frac{\tau}{\delta}\|u - v - \rho(Au - Av)\|. \end{aligned} \tag{4.7}$$

Since X is q-uniformly smooth, we have

$$\begin{aligned} \|u - v - \rho(Au - Av)\|^q &\leq \|u - v\|^q - q\rho\langle Au - Av, j_q(u - v)\rangle \\ &\quad + c_q\rho^q\|Au - Av\|^q \\ &\leq (1 - qr\rho + c_q s^q \rho^q)\|u - v\|^q. \end{aligned} \tag{4.8}$$

Combining (4.7) and (4.8), then we have

$$\|F(u) - F(v)\| \leq \frac{\tau}{\delta}(1 - qr\rho + c_q s^q \rho^q)^{\frac{1}{q}}\|u - v\|.$$

It follows from (4.5) that F is a contractive mapping and so there exists a unique point $u^* \in X$ such that

$$u^* = J_\rho^M(u^* - \rho Au^*).$$

Therefore, by Lemma 4.1, $u^* \in X$ is a unique solution of the problem (4.1). This completes the proof.

Remark 4.2. If X is 2-uniformly smooth and there exists $\rho > 0$ such that

$$|\rho - \frac{r}{c_2 s^2}| < \frac{\sqrt{r^2\tau^2 - c_2 s^2(\tau^2 - \delta^2)}}{c_2\tau s^2}, \quad r^2\tau^2 > c_2 s^2(\tau^2 - \delta^2), \tag{4.9}$$

then (4.5) holds. We note that all Hilbert spaces and L_p(or l_q) spaces, $2 \leq q < \infty$, are 2-uniformly smooth.

Algorithm 4.1. *Let $\eta : X \times X \to X^*$, $A : X \to X$ be two single-valued mappings and $M_n : X \to 2^X$ be a generalized m-accretive multivalued mapping for each $i = 1, 2, 3 \cdots$. For any given $u_0 \in X$, define an iterative sequence $\{u_n\}$ by*

$$\begin{cases} u_{n+1} = (1 - \alpha_n)u_n + \alpha_n J_\rho^{M_n}(v_n - \rho Av_n), \\ v_n = (1 - \beta_n)u_n + \beta_n J_\rho^{M_n}(u_n - \rho Au_n), \end{cases} \tag{4.10}$$

where $\{\alpha_n\}$ *and* $\{\beta_n\}$ *be two sequences in* $[0,1]$ *satisfying the following*

$$\sum_{n=0}^{\infty}\alpha_n = +\infty.$$

Theorem 4.2. *Let* X, η, A *and* M *be the same as in Theorem 4.1. Let* $M_n : X \to 2^X$ *be a generalized* m*-accretive multivalued mapping for each* $i = 1, 2, 3, \cdots$. *Assume that there exists* $\rho > 0$ *such that, for all* $z \in X$,

$$\lim_{n\to\infty} \|J_\rho^{M_n}(z) - J_\rho^M(z)\| = 0$$

and the condition (4.5) *holds. Then the iterative sequence* $\{u_n\}$ *generated by Algorithm 4.1 converges strongly to the unique solution* $u^* \in X$ *of the problem* (4.1).

Proof. Let $u^* \in X$ be the unique solution of the problem (4.1). It follows from Lemma 4.1 that

$$\begin{aligned} u^* &= (1-\alpha_n)u^* + \alpha_n J_\rho^M(u^* - \rho A u^*) \\ &= (1-\beta_n)u^* + \beta_n J_\rho^M(u^* - \rho A u^*). \end{aligned} \tag{4.11}$$

From (4.10) and (4.11), we have

$$\begin{aligned} &\|u_{n+1} - u^*\| \\ &\le (1-\alpha_n)\|u_n - u^*\| + \alpha_n \|J_\rho^{M_n}(v_n - \rho A v_n) - J_\rho^M(u^* - \rho A u^*)\| \\ &\le (1-\alpha_n)\|u_n - u^*\| + \alpha_n \|J_\rho^{M_n}(v_n - \rho A v_n) - J_\rho^{M_n}(u^* - \rho A u^*)\| \\ &\quad + \alpha_n \|J_\rho^{M_n}(u^* - \rho A u^*) - J_\rho^M(u^* - \rho A u^*)\| \\ &\le (1-\alpha_n)\|u_n - u^*\| + \alpha_n \frac{\tau}{\delta}\|v_n - u^* - \rho(Av_n - Au^*)\| + \alpha_n g_n, \end{aligned} \tag{4.12}$$

where

$$g_n = \|J_\rho^{M_n}(u^* - \rho A u^*) - J_\rho^M(u^* - \rho A u^*)\| \to 0.$$

Since X is q-uniformly smooth and A is strongly accretive and Lipschitz continuous, we have

$$\begin{aligned} &\|v_n - u^* - \rho(Av_n - Au^*)\|^q \\ &\le \|v_n - u^*\|^q - q\rho\langle Av_n - Au^*, j_q(v_n - u^*)\rangle + c_q\rho^q\|Av_n - Au^*\|^q \\ &\le (1 - qr\rho + c_q s^q \rho^q)\|v_n - u^*\|^q. \end{aligned} \tag{4.13}$$

Substituting (4.13) into (4.12), then we obtain

$$\|u_{n+1} - u^*\| \le (1-\alpha_n)\|u_n - u^*\| + \alpha_n h\|v_n - u^*\| + \alpha_n g_n, \tag{4.14}$$

where

$$h = \frac{\tau}{\delta}(1 - qr\rho + c_q s^q \rho^q)^{\frac{1}{q}} < 1.$$

Similarly, we can prove that

$$\|v_n - u^*\| \le (1-\beta_n)\|u_n - u^*\| + \beta_n h\|u_n - u^*\| + \alpha_n g_n. \tag{4.15}$$

It follows from (4.14) and (4.15) that

$$\begin{aligned} \|u_{n+1} - u^*\| &\le (1-\alpha_n)\|u_n - u^*\| + \alpha_n h(1-\beta_n)\|u_n - u^*\| \\ &\quad + \alpha_n\beta_n h^2\|u_n - u^*\| + \alpha_n^2 h g_n + \alpha_n g_n \\ &\le (1 - \alpha_n(1-h))\|u_n - u^*\| + \alpha_n g_n(\alpha_n h + 1). \end{aligned} \tag{4.16}$$

Letting

$$a_n = \|u_n - u^*\|,\ t_n = (1-h)\alpha_n,\ b_n = g_n\frac{(\alpha_n h+1)}{(1-h)},\ c_n = 0,$$

then (4.16) can be rewritten as

$$a_{n+1} \le (1-t_n)a_n + b_n t_n + c_n.$$

It follows from Lemma 4.2 that $a_n \to 0$ as $n \to \infty$ and so the sequence $\{u_n\}$ converges strongly to the unique solution $u^* \in X$ of the problem (4.1). This completes the proof.

Corollary 4.2. *Let X be a 2-uniformly smooth Banach space and η, A, M, M_n be the same as in Theorem 4.2. If the condition* (4.9) *holds, then the iterative sequence $\{u_n\}$ generated by Algorithm 4.1 converges strongly to the unique solution $u^* \in X$ of the problem* (4.1).

5. Acknowledgement

The first and second authors are supported by the National Natural Science Foundation of China and the Scientific Research Foundation for the Returned Overseas Chinese Scholars, State Education Ministry, and the third author is supported financially by the Korea Research Foundation Grant (KRF 2001-005-D00002).

References

[1] R. P. Agarwal, Y. J. Cho and N. J. Huang, *Sensitivity analysis for strongly nonlinear quasi-variational inclusions*, Appl. Math. Lett., **13(6)** (2000), 19–24.
[2] X. P. Ding and C. L. Luo, *Perturbed proximal point algorithms for generalized quasi-variational-like inclusions*, J. Comput. Appl. Math., **113** (2000), 153–165.
[3] A. Hassouni and A. Moudafi, *A perturbed algorithm for variational inequalities*, J. Math. Anal. Appl., **185** (1994), 706–712.
[4] N. J. Huang, *Generalized nonlinear variational inclusions with noncompact valued mapping*, Appl. Math. Lett., **9(3)** (1996), 25–29.
[5] N. J. Huang, *A new completely general class of variational inclusions with noncompact valued mappings*, Comput. Math. Appl., **35(6)** (1998), 9–14.
[6] N. J. Huang and C. X. Deng, *Auxiliary principle and iterative algorithms for generalized set-valued strongly nonlinear mixed variational-like inequalities*, J. Math. Anal. Appl., **256** (2001), 345–359.
[7] N. J. Huang and Y. P. Fang, *Generalized m-accretive mappings in Banach spaces*, J. Sichuan Univ., **38(4)** (2001), 591–592.
[8] K. R. Kazmi, *Mann and Ishikawa type perturbed iterative algorithms for generalized quasi-variational inclusions*, J. Math. Anal. Appl., **209** (1997), 572–584.
[9] C. H. Lee, Q. H. Ansari and J. C. Yao, *A perturbed algorithm for strongly nonlinear variational-like inclusions*, Bull. Austral. Math. Soc., **62** (2000), 417–426.
[10] M. V. Solodov and B. F. Svaiter, *A new projection method for variational inequality problems*, SIAM J. Control Optim., **37** (1999), 765–776.
[11] M. V. Solodov and P. Tseng, *Some methods based on the D-gap function for solving monotone variational inequalities*, Comput. Optim. Appl., **17** (2000), 255–277.
[12] P. Tseng, *Alternating projection-proximal methods for variational inequality problem*, SIAM J. Optim., **7** (1997), 951–965.
[13] R. U. Verma, *A class of projection-contraction methods applied to monotone variational inequalities*, Appl. Math. Lett., **13** (2000), 55–62.
[14] H. K. Xu, *Inequalities in Banach spaces with applications*, Nonlinear Anal., **16(12)** (1991), 1127–1138.

[15] George X. Z. Yuan, *KKM Theory and Applications in Nonlinear Analysis*, Marcel Dekker, New York, (1999).

JOURNAL OF CONCRETE AND APPLICABLE MATHEMATICS,VOL.3,NO.1,41-54,2005,COPYRIGHT 2005 EUDOXUS PRESS,LLC

$C^{1,1}$ functions and optimality conditions

Davide La Torre* Matteo Rocca†

Abstract

In this work we provide a characterization of $C^{1,1}$ functions on $\mathbb{R}^n$ (that is, differentiable with locally Lipschitz partial derivatives) by means of second directional divided differences. In particular, we prove that the class of $C^{1,1}$ functions is equivalent to the class of functions with bounded second directional divided differences. From this result we deduce a Taylor's formula for this class of functions and some optimality conditions. The characterizations and the optimality conditions proved by Riemann derivatives can be useful to write minimization algorithms; in fact, only the values of the function are required to compute second order conditions.

Keywords: Divided differences, Riemann derivatives, $C^{1,1}$ functions, nonlinear optimization, generalized derivatives

1 Introduction

The study of the class of $C^{1,1}$ functions has been renewed since the work of Hiriart-Urruty in his doctoral thesis [7]. The need for investigating these functions, as

*Contact author: University of Milan, Department of Economics, Faculty of Political Sciences, via Conservatorio,7, 20122, Milano, Italy. Tel.: +39276074462. Fax: +39276023198 E-mail: davide.latorre@unimi.it

†University of Insubria, Department of Economics, Faculty of Economics, via Ravasi, 2, 21100, Varese, Italy. Tel: +39258365693. Fax: +39258365617. E-mail: mrocca@eco.uninsubria.it

pointed out in [8], [10], [23], [24] and [25], comes from the fact that several problems of applied mathematics including variational inequalities, semi-infinite programming, iterated local minimization, etc. involve differentiable functions with no hope to be twice differentiable. In [8] the authors introduced the concept of generalized Hessian matrices and derived second order optimality conditions for nonlinear constrained problems. Further applications can be found in [10], [15], [19], [20], [22].
In this section we recall some concepts which are fundamental for understanding the proof of the results.

1.1 Riemann derivatives

In the following we will consider a function $f : \Omega \to \mathbb{R}$, with Ω an open subset of $\mathbb{R}^n$. For such a function we define:

$$\delta_2^d f(x;h) = f(x+2hd) - 2f(x+hd) + f(x).$$

with $x \in \Omega, h \in \mathbb{R}$ and $d \in \mathbb{R}^n$.

Definition 1.1. *The second Riemann derivative of f at a point $x \in \Omega$ in the direction $d \in \mathbb{R}^n$ is defined as:*

$$f_r''(x;d) = \lim_{h\to 0} \frac{\delta_2^d f(x;h)}{h^2},$$

if this limit exists.

Definition 1.2. *The second upper and lower Riemann derivatives of f at $x \in \Omega$ in the direction $d \in \mathbb{R}^n$ are defined, respectively, as:*

$$\overline{f}_r''(x;d) = \limsup_{h\to 0} \frac{\delta_2^d f(x;h)}{h^2},$$

$$\underline{f}_r''(x;d) = \liminf_{h\to 0} \frac{\delta_2^d f(x;h)}{h^2}.$$

Similarly we can define differences:

$$\Delta_2^d f(x;h) = f(x+hd) - 2f(x) + f(x-hd),$$

and then the corresponding second Riemann-type derivatives $f_R''(x;d)$, $\overline{f}_R''(x;d)$ and $\underline{f}_R''(x;d)$.

For properties of Riemann derivatives one can see [1], [2], [6] and [16].

Lemma 1.1. *Assume that f is bounded in a neighborhood of the point $x_0 \in \Omega$. If, for a fixed $d \in \mathbb{R}^n$, there exist neighborhoods U of the point x_0 and V of $0 \in \mathbb{R}$ such that $\frac{\delta_2^d f(x;h)}{h^2}$ is bounded on $U \times V \backslash \{0\}$, then also $\frac{f(x+hd)-f(x)}{h}$ is bounded on $U \times V \backslash \{0\}$.*

Proof. From the hypotheses we obtain that there exists a number $\delta > 0$ such that $\forall x \in U$ and $\forall h$ with $|h| \leq \delta, h \neq 0$, the following inequalities hold:

$$\left| f(x+hd) - f(x) - 2\left(f\left(x + \frac{h}{2}d\right) - f(x)\right)\right| \leq M \left|\frac{h}{2}\right|^2,$$

$$\left| f\left(x + \frac{h}{2}d\right) - f(x) - 2\left(f\left(x + \frac{h}{4}d\right) - f(x)\right)\right| \leq M \left|\frac{h}{4}\right|^2, \ldots$$

$$\left| f\left(x + \frac{h}{2^{n-1}}d\right) - f(x) - 2\left(f\left(x + \frac{h}{2^n}d\right) - f(x)\right)\right| \leq M \left|\frac{h}{2^n}\right|^2.$$

Multiplying these inequalities by $1, 2, 2^2, \ldots, 2^{(n-1)}$ respectively, we obtain by addition:

$$\left| f(x+hd) - f(x) - 2^n\left(f\left(x + \frac{h}{2^n}d\right) - f(x)\right)\right| \leq 2M \left|\frac{h}{2}\right|^2,$$

and hence:

$$\left| 2^n \frac{f(x + \frac{h}{2^n}d) - f(x)}{h}\right| \leq M'$$

for $\frac{1}{2}\delta \leq |h| \leq \delta$, by using the boundedness of f. Hence, writing $\xi = \frac{h}{2^n}$, we have:

$$\left| \frac{f(x+\xi d) - f(x)}{\xi}\right| \leq M' \quad for \ \frac{\delta}{2^{n+1}} \leq |\xi| \leq \frac{\delta}{2^n}, \ n = 0, 1, \ldots,$$

and the lemma is established, since n can be arbitrarily chosen.

□

In the following we set:

$$f'(x;d) = \lim_{h \to 0} \frac{f(x+hd) - f(x)}{h}, \quad f''(x;d) = \lim_{h \to 0} \frac{f'(x+hd;d) - f'(x;d)}{h},$$

if these limits exist.

1.2 Standard mollifiers

The function:

$$\phi(x) = \begin{cases} C\exp(\frac{1}{\|x\|^2-1}), & \text{if } \|x\| < 1 \\ \\ 0, & \text{if } \|x\| \geq 1 \end{cases}$$

is $C^\infty(\mathbb{R}^n)$ and we can choose the constant $C \in \mathbb{R}$ such that:

$$\int_{\mathbb{R}^n} \phi(x)dx = 1.$$

Definition 1.3. *Let $\varepsilon > 0$. The following functions:*

$$\phi_\varepsilon(x) = \frac{\phi(\frac{x}{\varepsilon})}{\varepsilon^n}$$

are called standard mollifiers.

Definition 1.4. *Let $f : \Omega \to \mathbb{R}$. We say that $f \in C_0^k(\Omega)$ if $f \in C^k(\Omega)$ and*

$$spt_f = \overline{\{x \in \Omega : f(x) \neq 0\}} \subset \Omega.$$

Theorem 1.1. *[3] The functions ϕ_ε are $C^\infty(\mathbb{R}^n)$ and satisfy:*

- $\int_{\mathbb{R}^n} \phi_\varepsilon(x)dx = 1$
- $spt_{\phi_\varepsilon} \subset B(0,\varepsilon) = \{x \in \mathbb{R}^n : \|x\| < \varepsilon\}$.

For a bounded function $f : \Omega \to \mathbb{R}$, and $\varepsilon > 0$, we define functions $f_\varepsilon : \mathbb{R}^n \to \mathbb{R}$ by the convolution $f_\varepsilon(x) = \int_\Omega \phi_\varepsilon(y-x)f(y)dy$. Observe that $f_\varepsilon(x) = 0$ if $x \notin \Omega + B(0,\varepsilon)$ and that $f_\varepsilon \in C^\infty(\mathbb{R}^n)$.

Theorem 1.2. *[3] Suppose that $f \in L^1_{loc}(\Omega)$. Then $f_\varepsilon(x) \to f(x)$ a.e. $x \in \Omega$, when $\varepsilon \to 0$. If $f \in C(\Omega)$ then the convergence is uniform on compact subsets of Ω.*

Theorem 1.3. *[3] Let K be a compact subset of Ω. Then $\exists \varepsilon_0 > 0$ such that $\forall \varepsilon \leq \varepsilon_0$ and $\forall x \in K$, the following function:*

$$y \to \phi_\varepsilon(y - x)$$

is $C_0^\infty(\Omega)$.

2 The main results

Definition 2.1. *A function $f : \Omega \to \mathbb{R}$ is locally Lipschitz at x_0 when there exist a constant K and a neighborhood U of x_0 such that:*

$$|f(x) - f(y)| \leq K\|x - y\|, \quad \forall x, y \in U.$$

Definition 2.2. *A function $f : \Omega \to \mathbb{R}$ is of class $C^{1,1}$ at x_0 when its first order partial derivatives exist in a neighborhood of x_0 and are locally Lipschitz at x_0.*

Some possible applications of $C^{1,1}$ functions are shown in the following examples.

Example 2.1. *Let $g : \Omega \subset \mathbb{R}^n \to \mathbb{R}$ be twice continuously differentiable on Ω and consider*[1] *$f(x) = [g^+(x)]^2$ where $g^+(x) = \max\{g(x), 0\}$. Then f is $C^{1,1}$ on Ω.*

Example 2.2. *In many problems in engineering applications and control theory ([23], [24] and the references therein) one has to study nonsmooth semi-infinite optimization problems such as the following:*

$$\textit{minimize } f(x)$$

$$\textit{subject to } \max_{t\in[a,b]} \phi_j(x,t) \leq 0, \ j = 1 \dots l$$

where $f : \mathbb{R}^n \to \mathbb{R}$ and $\phi_j : \mathbb{R}^n \to \mathbb{R}$ are C^2, $j = 1 \dots l$, $-\infty < a < x < b < +\infty$. One approach for solving this problem is to convert the functional constraints into equality constraints of the form:

$$h_j(x) = \int_a^b [\max\{\phi_j(x,y), 0\}]^2 \, dt = 0, j = 1 \dots l$$

and apply the methods of nonlinear programming. Hence the problem becomes:

$$\textit{minimize } f(x)$$

$$\textit{subject to } h_j(x) = 0, \ j = 1 \dots l$$

Since ϕ_j is C^2, it is easy see that the function h_j is $C^{1,1}$ with the gradient:

$$\nabla h_j(x) = 2 \int_a^b \max\{\phi_j(x,t), 0\} \nabla \phi_j(x,t) dt, \ j = 1 \dots l.$$

[1] This type of functions arises in some penalty methods.

Example 2.3. *Consider the following minimization problem:*

$$\min f_0(x)$$

over all $x \in \mathbb{R}^n$ *such that* $f_1(x) \leq 0, \ldots f_m(x) \leq 0$. *Letting* r *denote a positive parameter, the augmented Lagrangian* L_r *[21] is defined on* $\mathbb{R}^n \times \mathbb{R}^m$ *as*

$$L_r(x,y) = f_0(x) + \frac{1}{4r}\sum_{i=1}^{m}\{[y_i + 2rf_i(x)]^+\}^2 - y_i^2.$$

From the general theory of duality which yields L_r *as a particular Lagrangian, we know that* $L_r(x,\cdot)$ *is concave and also that* $L_r(\cdot,y)$ *is convex whenever the minimization problem is a convex minimization problem. Upon setting* $y = 0$ *in the previous expression, we observe that:*

$$L_r(x,0) = f_0(x) + r\sum_{i=1}^{m}[f_i^+(x)]^2$$

is the ordinary penalized version of the minimization problem. L_r *is differentiable everywhere on* $\mathbb{R}^n \times \mathbb{R}^m$ *with:*

$$\nabla_x L_r(x,y) = \nabla f_0(x) + \sum_{j=1}^{m}[y_j + 2rf_j(x)]^+\nabla f_j(x)$$

$$\frac{\partial L_r}{\partial y_i}(x,y) = \max\{f_i(x), \frac{-y_i}{2r}\}, i = 1\ldots m.$$

When the f_i *are* C^2 *on* $\mathbb{R}^n$, L_r *is* $C^{1,1}$ *on* $\mathbb{R}^{n+m}$. *The dual problem corresponding to* L_r *is by definition:*

$$\max g_r(y)$$

over $y \in \mathbb{R}^m$, *where* $g_r(y) = \inf_{x\in\mathbb{R}^n} L_r(x,y)$. *In the convex case with* $r > 0$, g_r *is again* $C^{1,1}$ *concave function with the following uniform Lipschitz property on* ∇g,

$$|\nabla g_r(y) - \nabla g_r(x)| \leq \frac{1}{2r}|y - y'|, \ \forall y, y' \in \mathbb{R}^m.$$

The following result characterizes a function of class $C^{1,1}$ by the boundness of second-order divided differences.

Theorem 2.1. *Assume that the function* $f : \Omega \to \mathbb{R}$ *is bounded on a neighborhood of the point* $x_0 \in \Omega$. *Then* f *is of class* $C^{1,1}$ *at* x_0 *if and only if there exist neighborhoods* U *of* x_0 *and* V *of* $0 \in \mathbb{R}$ *such that* $\frac{\delta_2^d f(x;h)}{h^2}$ *is bounded on* $U \times V\backslash\{0\}$, $\forall d \in S^1 = \{d \in \mathbb{R}^n : \|d\| = 1\}$.

Proof. i) Sufficiency. ¿From lemma 1.1, since $\frac{\delta_2^d f(x;h)}{h^2}$ is bounded on $U \times V \backslash \{0\}$, $\forall d \in S^1$, the same holds for $\frac{f(x+hd)-f(x)}{h}$. Observe that this last fact implies that f is locally Lipschitz at x_0 and hence continuous in a neighborhood of x_0. For every x in a neighborhood of x_0 and for ε "sufficiently small", we have, for $d \in S^1$:

$$f'_\varepsilon(x;d) = \lim_{h\to 0} \frac{f_\varepsilon(x+hd) - f_\varepsilon(x)}{h} =$$

$$\lim_{h\to 0} \frac{1}{h}\left[\int_\Omega \phi_\varepsilon(y-x-hd)f(y)dy - \int_\Omega \phi_\varepsilon(y-x)f(y)dy\right].$$

Putting $z = y - hd$, we obtain:

$$\int_\Omega \phi_\varepsilon(y-x-hd)f(y)dy = \int_{\Omega-\{hd\}} \phi_\varepsilon(z-x)f(z+hd)dz.$$

From theorem 1.3, we know that, for ε "sufficiently small", the functions $z \to \phi_\epsilon(z-x)$ are $C_0^\infty(\Omega)$ and hence, if also $|h|$ is "small enough", we get:

$$\int_{\Omega-\{hd\}} \phi_\varepsilon(z-x)f(z+hd)dz = \int_\Omega \phi_\varepsilon(z-x)f(z+hd)dz.$$

It follows that:

$$f'_\varepsilon(x;d) = \lim_{h\to 0} \int_\Omega \frac{f(z+hd)-f(z)}{h}\phi_\varepsilon(z-x)dz.$$

Furthermore one can easily see that:

$$f''_\varepsilon(x;d) = \lim_{h\to 0} \frac{f_\varepsilon(x+2hd) - 2f_\varepsilon(x+hd) + f_\varepsilon(x)}{h^2}$$

and similarly deduce:

$$f''_\varepsilon(x;d) = \lim_{h\to 0} \int_\Omega \frac{\delta_2^d f(z;h)}{h^2}\phi_\varepsilon(z-x)dz.$$

From the boundedness of $\frac{\delta_2^d f(x;h)}{h^2}$ and of $\frac{f(x+hd)-f(x)}{h}$, we obtain the existence of a constant M such that $|f'_\varepsilon(x;d)| \leq M$ and $|f''_\varepsilon(x;d)| \leq M$, for every $d \in S^1$, x in a neighborhood $\tilde{U}$ of x_0 and ε "sufficiently small". Hence, for every $x \in \tilde{U}$ and $d \in S^1$, there exists a sequence ε_n converging to 0 such that $f'_{\varepsilon_n}(x;d)$ converges to a limit which we denote by $\alpha(x;d)$. Observe that $\alpha(x;d)$ is bounded on $\tilde{U}$ whenever $d \in S^1$. For every $x \in \tilde{U},\ d \in S^1$ and h with $|h|$ "small enough", we can write:

$$f_{\varepsilon_n}(x+hd) = f_{\varepsilon_n}(x) + hf'_{\varepsilon_n}(x;d) + \frac{1}{2}h^2 f''_{\varepsilon_n}(\xi_n;d),$$

where $\xi_n \in (x, x+hd)$.
Recalling theorem 1.2, taking the limit for $n \to +\infty$, it follows that $f''_{\varepsilon_n}(\xi_n; d)$ converges to a limit which we denote by $\beta(x, h, d)$. Moreover:

$$f(x+hd) = f(x) + h\alpha(x; d) + \frac{1}{2}h^2\beta(x, h, d).$$

Observing that $\beta(x, h, d)$ is bounded for $x \in \tilde{U}$, $|h|$ "sufficiently small" and $d \in S^1$, it follows that $\alpha(x; d) = f'(x; d)$.
Furthermore, $\forall d \in S^1$ the functions $f''_{\varepsilon_n}(x; d)$ are bounded on $\tilde{U}$ uniformly with respect to ε and thus the functions $f'_{\varepsilon_n}(x; d)$ satisfy the following uniform Lipschitz condition:

$$\left|f'_{\varepsilon_n}(y; d) - f'_{\varepsilon_n}(x; d)\right| \leq B\|y - x\|, \forall x, y \in \tilde{U}.$$

Since $f'_{\varepsilon_n}(y; d)$ and $f'_{\varepsilon_n}(x; d)$ converge to $f'(y; d)$ and $f'(x; d)$ respectively, we see that $f'(x; d)$ is Lipschitz on $\tilde{U}$, $\forall d \in \mathbb{R}^n$. Taking $d = e^i, i = 1, \ldots, n$ (where e^i is the i-th fundamental vector of $\mathbb{R}^n$), we obtain the thesis.
ii) Necessity. Assume that f is of class $C^{1,1}$ at x_0. Set:

$$\overline{\Delta}_2^d f(x; s, t) = f(x + sd + td) - f(x + td) - f(x + sd) + f(x),$$

where $d \in S^1$, $x \in \Omega$, $s, t \in \mathbb{R}$ and $|s|$ and $|t|$ are "sufficiently small". Applying the mean value theorem, we obtain:

$$\frac{\overline{\Delta}_2^d f(x; s, t)}{st} = \frac{< \nabla f(x + \theta td + sd) - \nabla f(x + \theta td), d >}{s},$$

where $\theta \in (0, 1)$. Since f is of class $C^{1,1}$ at x_0 it follows easily that there exist a constant M, a neighborhood $\tilde{U}$ of x_0 and a number $\delta > 0$ such that, $\forall d \in S^1$ we have:

$$\left|\frac{\overline{\Delta}_2^d f(x; s, t)}{st}\right| \leq M, \quad \forall x \in \tilde{U}, \quad |s| < \delta, \ |t| < \delta.$$

Now the thesis follows observing that if $s = t = h$, then $\overline{\Delta}_2^d f(x; s, t) = \delta_2^d f(x; h)$.

□

Corollary 2.1. *Assume that the function f is bounded on a neighborhood of $x_0 \in \Omega$. Then f is of class $C^{1,1}$ at x_0 if and only if there exist neighborhoods U of x_0 and V of $0 \in \mathbb{R}$ such that $\frac{\Delta_2^d f(x;h)}{h^2}$ is bounded on $U \times V \backslash \{0\}$, $\forall d \in S^1$.*

Proof. The proof is straightforward remembering that:

$$\delta_2^d f(x;h) = \Delta_2^d f(x+hd;h).$$

□

Corollary 2.2. *If f is of class $C^{1,1}$ at x_0, there exist sequences ε_n converging to 0 and $\xi_n \in (x_0, x_0+hd)$ such that $f''_{\varepsilon_n}(\xi_n;d)$ converges to a limit $\beta(x_0,h,d)$ and it holds:*

$$f(x_0+hd) = f(x_0) + f'(x_0;d)h + \frac{\beta(x_0,h,d)}{2}h^2.$$

Proof. It is enclosed in the proof of the previous theorem. □

Theorem 2.2. *(Taylor's formula) Let f be a function of class $C^{1,1}$ at x_0.*

(i) If the function $x \to \overline{f}''_r(x;d)$ is upper semicontinuous in a neighborhood of x_0, for a fixed $d \in S^1$, then there exists $\xi \in [x_0, x_0+hd]$ such that, for h "small enough" we have:

$$f(x_0+hd) \le f(x_0) + hf'(x_0;d) + \frac{h^2}{2!}\overline{f}''_r(\xi;d)$$

(ii) If the function $x \to \underline{f}''_r(x;d)$ is lower semicontinuous in a neighborhood of x_0, for a fixed $d \in S^1$, then there exists $\xi \in [x_0, x_0+hd]$ such that for h "small enough" we have:

$$f(x_0+hd) \ge f(x_0) + hf'(x_0;d) + \frac{h^2}{2!}\underline{f}''_r(\xi;d)$$

Proof. i) Without loss of generality, the term $\beta(x_0;h;d)$ in the previous corollary 2.2 can be expressed as:

$$\beta(x_0;h;d) = \lim_{n\to+\infty} f''_{\epsilon_n}(\xi_n;d)$$

for some sequences $\xi_n \to \xi \in [x_0, x_0+hd]$ and $\epsilon_n \to 0$. Similarly to the proof of theorem 2.1, one can write that[2]:

$$f''_{\varepsilon_n}(\xi_n,d) = \int_\Omega \phi''_{\varepsilon_n}(y-\xi_n;d)f(y)dy =$$

$$\int_{B(0,1)} \lim_{h\to 0} \frac{\delta_2^d \phi_{\varepsilon_n}(y-\xi_n;h)}{h^2} f(y)dy =$$

[2]In the proof of this theorem we will use the following generalized version of Fatou's lemma: *if f_n is a sequence of measurable functions, $f_n \le M$ and $E \subset \mathbb{R}^n$ is a subset of finite measure, then* $\limsup_{n\to+\infty} \int_E f_n \le \int_E \limsup_{n\to+\infty} f_n$

$$\lim_{h\to 0}\int_{B(0,1)}\frac{\delta_2^d\phi_{\varepsilon_n}(y-\xi_n;h)}{h^2}f(y)dy=$$

$$\lim_{h\to 0}\int_{B(0,1)}\frac{\delta_2^d f(\xi_n+\varepsilon_n y;h)}{h^2}\phi_{\varepsilon_n}(y)dy\leq$$

$$\int_{B(0,1)}\limsup_{h\to 0}\frac{\delta_2^d f(\xi_n+\epsilon_n y;h)}{h^2}\phi(y)dy=\int_{B(0,1)}\overline{f}_r''(\xi_n+\epsilon_n y;d)\phi(y)dy.$$

Now using the upper semicontinuity of $\overline{f}_r''(\cdot;d)$ we have:

$$\beta(x;h;d)\leq\int_{B(0,1)}\limsup_{n\to+\infty}\overline{f}_r''(\xi_n+\epsilon_n y;d)\phi(y)dy\leq$$

$$\int_{B(0,1)}\overline{f}_r''(\xi;d)\phi(y)dy=\overline{f}_r''(\xi;d)$$

and the proof is complete.

ii) It is similar to the previous proof and we omit it. □

Theorem 2.3. *Assume that f is continuous and $f_r''(x;d)$ exists on a neighborhood of the point x_0 , $\forall d\in S^1$. Then f is of class $C^{1,1}$ at x_0 if and only if there exist a neighborhood U of x_0 and a function $g\in L^1(U)$ such that the following assumptions hold:*

(i) $\exists M\geq 0$ *such that* $|f_r''(x;d)|\leq M$, $\forall x\in U$, $\forall d\in S^1$,

(ii) $\left|\frac{\delta_2^d f(x;h)}{h^2}\right|\leq g(x)$, *for* $|h|$ *"small enough" (*$h\neq 0$*),* $d\in S^1$ *and a.e.* $x\in U$.

Proof. i) Sufficiency. Arguing in a fashion similar to that of theorem 2.1 and using Lebesgue theorem, we obtain for ε "sufficiently small", for every x in a neighborhood of x_0 and $d\in S^1$:

$$f_\varepsilon''(x;d)=\lim_{h\to 0}\int_\Omega\frac{\delta_2^d f(z;h)}{h^2}\phi_\varepsilon(z-x)dz=$$

$$\int_\Omega\lim_{h\to 0}\frac{\delta_2^d f(z;h)}{h^2}\phi_\varepsilon(z-x)dz=$$

$$\int_\Omega f_r''(z;d)\phi_\varepsilon(z-x)dz.$$

It follows that $\forall d\in S^1$ $f_\varepsilon''(x,d)$ is bounded on U (uniformly with respect to ε). Using the integral representation of divided differences (see for instance [9], ch. 6, th. 2), we have:

$$\frac{\delta_2^d f_\varepsilon(x;h)}{h^2}=\int_0^1 dt_1\int_0^{t_1}f_\varepsilon''(x+t_2hd+t_1hd;d)dt_2.$$

For x and h in suitable neighborhoods of x_0 and 0 respectively, the left member in the previous inequality is bounded by a constant M (uniformly with respect to ε). Sending ε to 0 and recalling theorem 1.2, we get the existence of neighborhoods U of x_0 and V of $0 \in \mathbb{R}$ such that $\forall d \in S^1$ $\frac{\delta_2^d f(x;h)}{h^2}$ is bounded on $U \times V \backslash \{0\}$. The thesis now follows recalling theorem 2.1.

ii) Necessity. The proof is similar to that of the necessary condition in theorem 2.1. □

Remark 2.1. *Hypothesis (ii) in the previous theorem cannot be omitted. In fact, as is easily seen, the function $f(x) = |x|$ satisfies hypothesis (i) but not hypothesis (ii) in a neighborhood of $x_0 = 0$ and is not of class $C^{1,1}$ at 0.*

Remark 2.2. *Theorems 2.1 and 2.3 extend the elementary condition which relates the Lipschitz condition on f' and the boundedness of f''. We generalize this relation without requiring any differentiability hypothesis and linking the existence and the Lipschitz behaviour of f' to the boundedness of $\frac{\delta_2^d f(x,h)}{h^2}$ or of the directional Riemann derivatives.*

Remark 2.3. *Conditions similar to those of theorem 2.3, expressed in terms of $f''_R(x;d)$ can be proved in analogous way.*

3 Optimality conditions for unconstrained optimization problems

The aim of this section is to study necessary and sufficient conditions for $C^{1,1}$ unconstrained optimization problems. These conditions are proved by using the generalized Taylor's expansions given in the previous section and Riemann derivatives. These optimality conditions can be used to write minimization algorithms; in fact for the computation of the next results only the values of the function are required. In the following we will suppose that, for any $d \in S^1$, the function $x \to \overline{f}''_r(x,d)$ is upper semicontinuous and that the function $x \to \underline{f}''_r(x;d)$ is lower semicontinuous in a neighborhood of x_0.

So we consider the following unconstrained optimization problem:

$P1)$ $$\min_{x \in \Omega} f(x)$$

where Ω is an open subset of $\mathbb{R}^n$.

Theorem 3.1. *(necessary condition) If f is $C^{1,1}$ and x_0 is a local minimum point then $\nabla f(x_0) = 0$ and $\overline{f}''_r(x_0, d) \geq 0 \ \forall d \in S^1$.*

Proof. ¿From Taylor's formula we obtain, for h "small enough":

$$f(x_0 + hd) \leq f(x_0) + h < \nabla f(x_0), d > + \frac{h^2}{2} \overline{f}''_r(\xi, d)$$

where $\xi \in [x_0, x_0 + hd]$. So

$$0 \leq f(x_0 + hd) - f(x_0) \leq \overline{f}''_r(\xi, d)$$

and taking the limit when $h \to 0$ we obtain the thesis. □

Theorem 3.2. *(sufficient condition) If f is $C^{1,1}$, $\nabla f(x_0) = 0$ and $\underline{f}''_r(x_0 + \alpha d, d) > 0$, $\forall \alpha \in (0, 1)$ and $\forall d \in S^1$, then x_0 is a strict local minimum of f on Ω.*

Proof. On the contrary suppose that x_0 is not a strict local minimum; then there exists a sequence x_k such that $x_k \to x_0$, when $k \to +\infty$ and $f(x_k) \leq f(x_0) \ \forall k \in \mathbb{N}$. So $x_k = x_0 + \delta_k u_k$, where $\|u_k\| = 1$ and $\delta_k \to 0$ when $k \to +\infty$. So we have:

$$f(x_k) \geq f(x_0) + \delta_k^2 \frac{\underline{f}''_r(\xi_k, u_k)}{2}$$

where $x_0 \leq \xi_k \leq x_0 + \delta_k u_k$. This implies that

$$0 \geq f(x_k) - f(x_0) \geq \delta_k^2 \frac{\underline{f}''_r(\xi_k, u_k)}{2}$$

and then, when k is "sufficiently large", we obtain:

$$\underline{f}''_r(\xi_k, u_k) \leq 0$$

which contradicts the hypothesis. □

References

[1] Ash J.M.: *Very generalized Riemann derivatives.* Real Analysis Exhange, 12, 1985, 10-29.

[2] Ash J.M.: *Generalizations of the Riemann derivative.*, Tran. Amer. Math. Soc., 126, 1997, 181-199.

[3] H. Brezis: *Analyse fonctionelle- Theorie et applications.* Masson Editeur, Paris, 1963.

[4] X. Chen: *Convergence of BFGS method for LC^1 convex constrained optimization.* SIAM J. Control Optim., 34, 1996, 2051-2063.

[5] R. Cominetti, R. Correa: *A generalized second-order derivative in nonsmooth optimization.* SIAM J. Control Optim., 28, 1990, 789-809.

[6] A. Guerraggio, M. Rocca: *Derivate dirette di Riemann e di Peano.* Convessitá e Calcolo Parallelo, Verona, 1997.

[7] J.B. Hiriart-Hurruty: *Contributions a la programmation mathematique: deterministe et stocastique.* (Doctoral thesis), Univ. Clermont-Ferrand, 1977.

[8] J.B. Hiriart-Urruty, J.J. Strodiot , V. Hien Nguyen: *Generalized Hessian matrix and second order optimality conditions for problems with $C^{1,1}$ data.* Appl. Math. Optim., 11, 1984, 43-56.

[9] E. Isaacson, B.H. Keller: *Analysis of numerical methods.* Wiley, New York, 1966.

[10] D. Klatte, K. Tammer: *On the second order sufficient conditions for $C^{1,1}$-optimization problems.*, Optimization, 19, 1988, 169-180.

[11] D. Klatte: *Upper Lipschitz behavior of solutions to perturbed $C^{1,1}$ programs.* Math. Program. (Ser. B), 88, 2000, 285-311.

[12] D. La Torre, M. Rocca: *$C^{k,1}$ functions and Riemann derivatives.* Real Analysis Exchange, 25, 1999-2000,743-752.

[13] D. La Torre, M. Rocca: *Higher order smoothness conditions and differentiability of real functions.* Real Analysis Exchange, 26, 2000-01, 657-667.

[14] D. La Torre: *Characterizations of $C^{1,1}$ functions, Taylor's formulae and optimality conditions.* (Doctoral Thesis), Department of Mathematics, University of Milan, 2001.

[15] D.T. Luc: *Taylor's formula for $C^{k,1}$ functions.* SIAM J. Optimization, 5, 1995, 659-669.

[16] Marcinkiewicz J., Zygmund A.: *On the differentiability of functions and summability of trigonometrical series.* Fund. Math., 26, 1936, 1-43.

[17] P. Michel, J.P. Penot: *Second-order moderate derivatives.* Nonlin. Anal., 22, 1994, 809-824.

[18] L. Qi, W. Sun: *A nonsmooth version of Newton's method.* Math. Program., 58, 1993, 353-367.

[19] L. Qi: *Superlinearly convergent approximate Newton methods for LC^1 optimization problems.* Math. Program., 64, 1994, 277-294.

[20] L. Qi: *LC^1 functions and LC^1 optimization.* Operations Research and its applications (D.Z. Du, X.S. Zhang and K. Cheng eds.), World Publishing, Beijing, 1996, 4-13.

[21] T.D. Rockafellar: *Proximal subgradients, marginal values, and augmented lagrangians in nonconvex optimization.* Mathematics of Operations Research, 6, 1981, 427-437.

[22] W. Sun, R.J.B. de Sampaio, J. Yuan: *Two algorithms for LC^1 unconstrained optimization.* J. Comput. Math., 18, 2000, 621-632.

[23] X.Q. Yang, V. Jeyakumar: *Generalized second-order directional derivatives and optimization with $C^{1,1}$ functions.* Optimization, 26, 1992, 165-185.

[24] X.Q. Yang: *Second-order conditions in $C^{1,1}$ optimization with applications.* Numer. Funct. Anal. and Optimiz., 14, 1993, 621-632.

[25] X.Q. Yang: *Generalized second-order derivatives and optimality conditions.* Nonlin. Anal., 23, 1994, 767-784.

Numerical analysis of a quasistatic sliding contact problem with wear

José R. Fernández[(1)], Mircea Sofonea[(2)] and Juan M. Viaño[(1)]

(1) Departamento de Matemática Aplicada, Universidade de Santiago de Compostela, Campus Sur s/n, Facultade de Matemáticas. 15782 Santiago de Compostela, Spain
E-mail: jramon@usc.es, maviano@usc.es

(2) Laboratoire de Mathématiques et Physique pour les Systèmes (MEPS), Université de Perpignan via Domitia, 52 Avenue Paul Alduy, 66860 Perpignan, France,
E-mail: sofonea@univ-perp.fr

Abstract

We consider a mathematical model which describes the sliding contact with wear between a viscoelastic body and a rigid moving foundation. The process is quasistatic, the contact is bilateral and the wear is modeled with a version of Archard's law. The variational formulation of the problem consists of a system of nonlinear evolutionary equations which has a unique solution under certain assumptions on the given data. We study the numerical approach to the problem using a fully discrete finite elements scheme with implicit discretization in time and we derive error estimates for the approximative solutions. Finally, we present numerical results in the study of one and two dimmensional test problems.

Keywords: Viscoelasticity, contact, Archard's law, wear, finite elements.

1. Introduction

Wear is one of the plagues which reduce the lifetime of modern machine elements. It represents the unwanted removal of materials from surfaces of contacting bodies occuring in relative motion. Wear arises when a hard rough surface slides against a softer surface, digs into it, and its asperities plough a series of grooves. Generally, a mathematical theory of friction and wear should be a generalization of experimental facts and it must be in agreement with the laws of thermodynamics of irreversible processes. A general model of quasistatic frictional contact with wear between deformable bodies was derived [11], [12] from thermodynamic considerations. This model was used in various papers (see for example [9, 10]), where existence and uniqueness results of weak solutions have been proved.

The present paper is devoted to the study of the numerical analysis of a problem of sliding contact with wear. The process is quasistatic, and the contact is bilateral. We model the process as in [12] by introducing the wear function which measures the wear of the contact surface and which satisfies Archard's law. The variational analysis of the model was provided in [2]. Here we study the numerical approach to the problem using spatially semi-discrete and fully discrete finite elements schemes with implicit discretization in time. Notice that from physical point of view our model is quite limited and, since the processes involved in wear of contact surfaces are very complicated, there is a need to consider in the future more sophisticated mathematical models. However, we believe that the rigorous numerical analysis in the context of this simplified physical model may be helpful and represents a first step in the study of more realistic models describing the contact with wear.

The paper is organized as follows. In Section 2 we introduce some notations and preliminaries. In Section 3 we present the sliding contact problem and derive the variational formulation. Then, we recall a version of the result obtained in [2] concerning the unique solvability of the model. In Section 4 we consider the fully discrete approximation of the model, by using finite elements for the spatial variable and the backward Euler scheme for the time variable. We state the existence and uniqueness of the fully discrete solution and we derive error estimates which prove its convergence to the solution of the continuous problem. Finally, in Section 5 we present some numerical results for one and two dimensional test examples.

2. Notation and preliminaries

In this short section we present the notation we shall use and some preliminary material. For further details we refer the reader to [3, 6, 8].

We denote by $\mathbb{S}_d$ the space of second order symmetric tensors on $\mathbb{R}^d$ $(d = 1, 2, 3)$, while "$\cdot$" and $|\cdot|$ will represent the inner product and the Euclidean norm on $\mathbb{S}_d$ and $\mathbb{R}^d$, respectively, i.e.

$$\boldsymbol{u} \cdot \boldsymbol{v} = u_i v_i, \quad |\boldsymbol{v}| = (\boldsymbol{v}.\boldsymbol{v})^{\frac{1}{2}} \qquad \forall \boldsymbol{u}, \boldsymbol{v} \in \mathbb{R}^d,$$

$$\boldsymbol{\sigma} \cdot \boldsymbol{\tau} = \sigma_{ij}\tau_{ij}, \quad |\boldsymbol{\tau}| = (\boldsymbol{\tau} \cdot \boldsymbol{\tau})^{\frac{1}{2}} \qquad \forall \boldsymbol{\sigma}, \boldsymbol{\tau} \in \mathbb{S}_d.$$

Here and below the indices i and j run between 1 and d, and the summation convention over repeated indices is adopted.

Let $\Omega \subset \mathbb{R}^d$ be a bounded domain with a Lipschitz boundary Γ and let $\boldsymbol{\nu}$ denote the unit outer normal on Γ. We shall use the notation

$$H = [L^2(\Omega)]^d = \{\, \boldsymbol{u} = (u_i) \;;\; u_i \in L^2(\Omega) \,\}, \quad Q = \{\, \boldsymbol{\sigma} = (\sigma_{ij}) \;;\; \sigma_{ij} = \sigma_{ji} \in L^2(\Omega) \,\},$$

$$H_1 = \{\, \boldsymbol{u} = (u_i) \;;\; \boldsymbol{\varepsilon}(\boldsymbol{u}) \in Q \,\}, \quad Q_1 = \{\, \boldsymbol{\sigma} \in Q \;;\; \mathrm{Div}\, \boldsymbol{\sigma} \in H \,\}.$$

Here $\boldsymbol{\varepsilon} : H_1 \longrightarrow Q$ and $\mathrm{Div} : Q_1 \longrightarrow H$ are the *deformation* and the *divergence* operators, respectively, defined by

$$\boldsymbol{\varepsilon}(\boldsymbol{u}) = (\varepsilon_{ij}(\boldsymbol{u})), \quad \varepsilon_{ij}(\boldsymbol{u}) = \frac{1}{2}(u_{i,j} + u_{j,i}), \quad \mathrm{Div}\, \boldsymbol{\sigma} = (\sigma_{ij,j}),$$

where the index that follows a comma indicates a partial derivative with respect to the corresponding component of the independent variable. The spaces H, Q, H_1 and Q_1 are real Hilbert spaces endowed with the cannonical inner products given by

$$(\boldsymbol{u}, \boldsymbol{v})_H = \int_\Omega u_i v_i \, dx, \qquad (\boldsymbol{\sigma}, \boldsymbol{\tau})_Q = \int_\Omega \sigma_{ij} \tau_{ij} \, dx,$$

$$(\boldsymbol{u}, \boldsymbol{v})_{H_1} = (\boldsymbol{u}, \boldsymbol{v})_H + (\boldsymbol{\varepsilon}(\boldsymbol{u}), \boldsymbol{\varepsilon}(\boldsymbol{v}))_Q, \qquad (\boldsymbol{\sigma}, \boldsymbol{\tau})_{Q_1} = (\boldsymbol{\sigma}, \boldsymbol{\tau})_Q + (\mathrm{Div}\, \boldsymbol{\sigma}, \mathrm{Div}\, \boldsymbol{\tau})_H.$$

The associated norms on these spaces are denoted by $|\cdot|_H$, $|\cdot|_Q$, $|\cdot|_{H_1}$ and $|\cdot|_{Q_1}$, respectively.

Let $H_\Gamma = [H^{\frac{1}{2}}(\Gamma)]^d$ and let $\gamma : H_1 \longrightarrow H_\Gamma$ be the trace map. For every element $\boldsymbol{v} \in H_1$ we still write $\boldsymbol{v}$ to denote the trace $\gamma \boldsymbol{v}$ of $\boldsymbol{v}$ on Γ and we denote by v_ν and $\boldsymbol{v}_\tau$ the *normal* and the *tangential* components of $\boldsymbol{v}$ on the boundary Γ given by

$$v_\nu = \boldsymbol{v} \cdot \boldsymbol{\nu}, \quad \boldsymbol{v}_\tau = \boldsymbol{v} - v_\nu \boldsymbol{\nu}. \tag{2.1}$$

Let H_Γ' be the dual of H_Γ and let $(\cdot, \cdot)$ denote the duality pairing between H_Γ' and H_Γ. For every $\boldsymbol{\sigma} \in Q_1$ there exists an element $\boldsymbol{\sigma\nu} \in H_\Gamma'$ such that

$$(\boldsymbol{\sigma}, \boldsymbol{\varepsilon}(\boldsymbol{v}))_Q + (\mathrm{Div}\, \boldsymbol{\sigma},\, \boldsymbol{v})_H = (\boldsymbol{\sigma\nu}, \gamma \boldsymbol{v}) \qquad \forall \boldsymbol{v} \in H_1. \tag{2.2}$$

Moreover, if $\boldsymbol{\sigma}$ is a regular function, then

$$(\boldsymbol{\sigma\nu}, \gamma \boldsymbol{v}) = \int_\Gamma \boldsymbol{\sigma\nu} \cdot \boldsymbol{v} \, da \qquad \forall \boldsymbol{v} \in H_1. \tag{2.3}$$

We also denote by σ_ν and $\boldsymbol{\sigma}_\tau$ the *normal* and *tangential* traces of $\boldsymbol{\sigma}$ and we recall that, when $\boldsymbol{\sigma}$ is a regular (say C^1) function, then

$$\sigma_\nu = (\boldsymbol{\sigma\nu}) \cdot \boldsymbol{\nu}, \quad \boldsymbol{\sigma}_\tau = \boldsymbol{\sigma\nu} - \sigma_\nu \boldsymbol{\nu}. \tag{2.4}$$

Finally, for every real Hilbert space X, we use the classical notation for the space $L^\infty(0, T; X)$ and we denote by $C([0, T]; X)$ and $C^1([0, T]; X)$, $T > 0$, the space of continuous and continuously differentiable functions from $[0, T]$ to X, respectively.

3. The model of sliding contact with wear

In this section we describe the model for the process, present its variational formulation and state an existence and uniqueness result, Theorem 3.1.

The physical setting is as follows. A viscoelastic body occupies in its reference configuration a domain Ω with outer Lipschitz surface Γ that is divided into three disjoint measurable parts Γ_1, Γ_2 and Γ_3, such that $meas\,\Gamma_1 > 0$. Let $[0,T]$ be the time interval of interest, where $T > 0$. The body is clamped on $\Gamma_1 \times (0,T)$ and therefore the displacement field vanishes there. A volume force of density $\boldsymbol{f}_0$ acts in $\Omega \times (0,T)$ and surface tractions of density $\boldsymbol{f}_2$ act on $\Gamma_2 \times (0,T)$. On Γ_3 the body is in bilateral contact with a moving rigid foundation, which results in the wear of the contacting surface. We assume that there is only sliding contact, which is always maintained.

We denote by $\boldsymbol{u}$ the displacement field, $\boldsymbol{\sigma}$ the stress field and $\boldsymbol{\varepsilon}(\boldsymbol{u})$ the small strain tensor. We assume that the constitutive law of the material is of the form

$$\boldsymbol{\sigma} = \mathcal{A}\boldsymbol{\varepsilon}(\dot{\boldsymbol{u}}) + \mathcal{G}\boldsymbol{\varepsilon}(\boldsymbol{u}), \tag{3.1}$$

where $\mathcal{A}$ and $\mathcal{G}$ are given nonlinear functions which will be described below. In (3.1) and below, in order to simplify the notation, we usually do not indicate explicitely the dependence of various functions on the variables $\boldsymbol{x} \in \Omega \cup \Gamma$ and $t \in [0,T]$. Moreover, the dot above represents the derivative with respect to the time variable, i.e.

$$\dot{\boldsymbol{u}} = \frac{d\boldsymbol{u}}{dt}.$$

We recall that in linear viscoelasticity, the stress tensor $\boldsymbol{\sigma} = (\sigma_{ij})$ is given by

$$\sigma_{ij} = a_{ijkl}\varepsilon_{kl}(\dot{\boldsymbol{u}}) + g_{ijkl}\varepsilon_{kl}(\boldsymbol{u}), \tag{3.2}$$

where $\mathcal{A} = (a_{ijkl})$ is the viscosity tensor and $\mathcal{G} = (g_{ijkl})$ is the elasticity tensor, for $i,j,k,l = 1,...,d$. Kelvin-Voigt viscoelastic materials of the form (3.1) involving nonlinear constitutive functions have been considered recently in [9, 10].

Next we describe the conditions on the contact surface Γ_3 in which our interest lies. We denote by $\boldsymbol{v}^\star$ the velocity of the foundation and we assume that the tangential displacement on the contact surface vanishes, i.e.

$$\boldsymbol{u}_\tau = \boldsymbol{0}.$$

We introduce the wear function $w : \Gamma_3 \times [0,T] \longrightarrow \mathbb{R}$, which measures the wear of the surface, using the model derived in [11, 12]. The wear is identified as an increase in gap in the normal direction between the body and the foundation or, equivalently,

as the depth of the material that is lost. Since the body is in bilateral contact with the foundation it follows that

$$w = -u_\nu + u_{0\nu} \tag{3.3}$$

where $\boldsymbol{u}_0$ represents the initial displacement field. Thus, the position of the contact evolves with the wear. We remark that $w(0) = 0$ i.e. at the initial moment the material is new. Moreover, the effect of the wear is the recession of Γ_3 and therefore it is expected that $u_\nu \leq u_{0\nu}$, which implies $w \geq 0$. We conclude that the wear is non-negative which justifies the sign convention in (3.3).

The evolution of the wear of the contacting surface is governed by a version of Archard's law (see [11, 12]), which states that the rate of the wear is proportional to the contact stress and the slip, thus

$$\dot{w} = -k_w \sigma_\nu v^\star, \tag{3.4}$$

where $k_w > 0$ is a wear coefficient and $v^\star = |\boldsymbol{v}^\star| > 0$.

We can now eliminate the unknown function w from our problem. Let $\beta = 1/(k_w v^\star)$. Using (3.3) and (3.4), we have

$$\sigma_\nu = \beta \dot{u}_\nu. \tag{3.5}$$

Moreover, the wear increases in time, i.e. $\dot{w} \geq 0$ and therefore, from (3.3), it follows that $\dot{u}_\nu \leq 0$. Thus, the contact condition (3.5) implies

$$\sigma_\nu = -\beta |\dot{u}_\nu|.$$

Finally, we assume that the process is quasistatic, i.e. the inertial term can be neglected in the equation of motion.

With these assumptions, the classical formulation of the quasistatic problem of sliding contact with wear is the following one.

Problem P. Find a displacement field $\boldsymbol{u} : \Omega \times [0,T] \longrightarrow \mathbb{R}^d$ and a stress field $\boldsymbol{\sigma} : \Omega \times [0,T] \longrightarrow \mathbb{S}_d$ such that

$$\boldsymbol{\sigma} = \mathcal{A}\boldsymbol{\varepsilon}(\dot{\boldsymbol{u}}) + \mathcal{G}\boldsymbol{\varepsilon}(\boldsymbol{u}) \quad \text{in } \Omega \times (0,T), \tag{3.6}$$

$$\operatorname{Div} \boldsymbol{\sigma} + \boldsymbol{f}_0 = \boldsymbol{0} \quad \text{in } \Omega \times (0,T), \tag{3.7}$$

$$\boldsymbol{u} = \boldsymbol{0} \quad \text{on } \Gamma_1 \times (0,T), \tag{3.8}$$

$$\boldsymbol{\sigma\nu} = \boldsymbol{f}_2 \quad \text{on } \Gamma_2 \times (0,T), \tag{3.9}$$

$$\sigma_\nu = -\beta |\dot{u}_\nu|, \quad \boldsymbol{u}_\tau = \boldsymbol{0} \quad \text{on } \Gamma_3 \times (0,T), \tag{3.10}$$

$$\boldsymbol{u}(0) = \boldsymbol{u}_0 \quad \text{in } \Omega. \tag{3.11}$$

60 To obtain the variational formulation of Problem P, we need additional notation. To this end, let V denote the closed subspace of H_1, defined by

$$V = \{\boldsymbol{v} \in H_1 \ ; \ \boldsymbol{v} = \boldsymbol{0} \ \text{ on } \Gamma_1, \ \boldsymbol{v}_\tau = \boldsymbol{0} \ \text{ on } \Gamma_3\}.$$

Since $meas\,\Gamma_1 > 0$, Korn's inequality holds: there exists $C_K > 0$ which depends only on Ω and Γ_1 such that

$$|\boldsymbol{\varepsilon}(\boldsymbol{v})|_Q \geq C_K |\boldsymbol{v}|_{H_1} \qquad \forall \boldsymbol{v} \in V. \tag{3.12}$$

A proof of Korn's inequality may be found in [7, p. 79]. On V, we consider the inner product given by

$$(\boldsymbol{u}, \boldsymbol{v})_V = (\boldsymbol{\varepsilon}(\boldsymbol{u}), \boldsymbol{\varepsilon}(\boldsymbol{v}))_Q \qquad \forall \boldsymbol{u}, \ \boldsymbol{v} \in V,$$

and let $|\cdot|_V$ be the associated norm, i.e.

$$|\boldsymbol{v}|_V = |\boldsymbol{\varepsilon}(\boldsymbol{v})|_Q \qquad \forall \boldsymbol{v} \in V. \tag{3.13}$$

It follows from (3.12) and (3.13) that $|\cdot|_{H_1}$ and $|\cdot|_V$ are equivalent norms on V and therefore $(V, |\cdot|_V)$ is a real Hilbert space. Moreover, by the Sobolev's trace theorem and (3.12) we have a constant C_0 depending only on the domain Ω, Γ_1 and Γ_3 such that

$$|\boldsymbol{v}|_{[L^2(\Gamma_3)]^d} \leq C_0 |\boldsymbol{v}|_V \qquad \forall \boldsymbol{v} \in V. \tag{3.14}$$

In the study of the frictional problems with wear, we assume that the viscosity and elasticity operators satisfy

$$\left.\begin{array}{l}
\text{(a) } \mathcal{A} : \Omega \times \mathbb{S}_d \to \mathbb{S}_d. \\
\text{(b) There exists } \mathcal{L}_{\mathcal{A}} > 0 \text{ such that} \\
\qquad |\mathcal{A}(\boldsymbol{x}, \boldsymbol{\varepsilon}_1) - \mathcal{A}(\boldsymbol{x}, \boldsymbol{\varepsilon}_2)| \leq \mathcal{L}_{\mathcal{A}} |\boldsymbol{\varepsilon}_1 - \boldsymbol{\varepsilon}_2| \\
\qquad \forall\, \boldsymbol{\varepsilon}_1, \boldsymbol{\varepsilon}_2 \in \mathbb{S}_d, \ \text{a.e. } \boldsymbol{x} \in \Omega. \\
\text{(c) There exists } m_{\mathcal{A}} > 0 \text{ such that} \\
\qquad (\mathcal{A}(\boldsymbol{x}, \boldsymbol{\varepsilon}_1) - \mathcal{A}(\boldsymbol{x}, \boldsymbol{\varepsilon}_2)) \cdot (\boldsymbol{\varepsilon}_1 - \boldsymbol{\varepsilon}_2) \geq m_{\mathcal{A}} \, |\boldsymbol{\varepsilon}_1 - \boldsymbol{\varepsilon}_2|^2 \\
\qquad \forall\, \boldsymbol{\varepsilon}_1, \boldsymbol{\varepsilon}_2 \in \mathbb{S}_d, \ \text{a.e. } \boldsymbol{x} \in \Omega. \\
\text{(d) For any } \boldsymbol{\varepsilon} \in \mathbb{S}_d, \ \boldsymbol{x} \mapsto \mathcal{A}(\boldsymbol{x}, \boldsymbol{\varepsilon}) \text{ is Lebesgue measurable on } \Omega. \\
\text{(e) The mapping } \boldsymbol{x} \mapsto \mathcal{A}(\boldsymbol{x}, \boldsymbol{0}) \in Q.
\end{array}\right\} \tag{3.15}$$

$$\left.\begin{array}{l}
\text{(a) } \mathcal{G} : \Omega \times \mathbb{S}_d \to \mathbb{S}_d. \\
\text{(b) There exists an } \mathcal{L}_{\mathcal{G}} > 0 \text{ such that} \\
\qquad |\mathcal{G}(\boldsymbol{x}, \boldsymbol{\varepsilon}_1) - \mathcal{G}(\boldsymbol{x}, \boldsymbol{\varepsilon}_2)| \leq \mathcal{L}_{\mathcal{G}} \, |\boldsymbol{\varepsilon}_1 - \boldsymbol{\varepsilon}_2| \\
\qquad \forall\, \boldsymbol{\varepsilon}_1, \boldsymbol{\varepsilon}_2 \in \mathbb{S}_d, \ \text{a.e. } \boldsymbol{x} \in \Omega. \\
\text{(c) For any } \boldsymbol{\varepsilon} \in \mathbb{S}_d, \ \boldsymbol{x} \mapsto \mathcal{G}(\boldsymbol{x}, \boldsymbol{\varepsilon}) \text{ is Lebesgue measurable on } \Omega. \\
\text{(d) The mapping } \boldsymbol{x} \mapsto \mathcal{G}(\boldsymbol{x}, \boldsymbol{0}) \in Q.
\end{array}\right\} \tag{3.16}$$

We note that assumptions (3.15) and (3.16) are satisfied in the case of the linear viscoelastic model (3.2) provided that $a_{ijkl} \in L^\infty(\Omega)$, $g_{ijkl} \in L^\infty(\Omega)$ and there exists $\alpha > 0$ such that

$$a_{ijkl}(\boldsymbol{x})\,\xi_k\,\xi_l \geq \alpha\,|\boldsymbol{\xi}|^2 \qquad \forall\,\boldsymbol{\xi} \in \mathbb{S}_d,\ \text{a.e. } \boldsymbol{x} \in \Omega.$$

We also assume that the body forces and tractions have the regularity

$$\boldsymbol{f}_0 \in C([0,T];H), \quad \boldsymbol{f}_2 \in C([0,T];[L^2(\Gamma_2)]^d), \tag{3.17}$$

and the function β is such that

$$\beta \in L^\infty(\Gamma_3) \text{ and there exists } \beta_\star \text{ such that } \beta(\boldsymbol{x}) \geq \beta_\star > 0,\ \text{a.e. } \boldsymbol{x} \in \Gamma_3. \tag{3.18}$$

Finally, the initial displacement satisfies

$$\boldsymbol{u}_0 \ \in V. \tag{3.19}$$

Next, we define the element $\boldsymbol{f}(t) \in V$ given by

$$(\boldsymbol{f}(t),\boldsymbol{v})_V = (\boldsymbol{f}_0(t),\boldsymbol{v})_H + (\boldsymbol{f}_2(t),\boldsymbol{v})_{[L^2(\Gamma_2)]^d} \qquad \forall \boldsymbol{v} \in V, \quad t \in [0,T], \tag{3.20}$$

and let $j : V \times V \to \mathbb{R}$ be the functional

$$j(\boldsymbol{u},\boldsymbol{v}) = \int_{\Gamma_3} \beta|u_\nu|\,v_\nu\,da \qquad \forall\,\boldsymbol{u},\,\boldsymbol{v} \in V. \tag{3.21}$$

Keeping in mind (3.17) and (3.18) we note that the integrals in (3.20) and (3.21) are well defined. Moreover, using (3.20) it follows that

$$\boldsymbol{f} \in C([0,T];V). \tag{3.22}$$

Using (2.1)–(2.3) and (2.4), it is straightforward to show that if $\{\boldsymbol{u},\,\boldsymbol{\sigma}\}$ represents a regular solution to Problem P then

$$(\boldsymbol{\sigma}(t),\boldsymbol{\varepsilon}(\boldsymbol{v}))_Q + j(\dot{\boldsymbol{u}}(t),\boldsymbol{v}) = (\boldsymbol{f}(t),\boldsymbol{v})_V \qquad \forall \boldsymbol{v} \in V,\ t \in [0,T]. \tag{3.23}$$

To conclude, from (3.6), (3.11) and (3.23) we obtain the following variational formulation of the mechanical problem P.

Problem PV**.** *Find a displacement field* $\boldsymbol{u} : [0,T] \to V$ *and a stress field* $\boldsymbol{\sigma} : [0,T] \to Q$ *such that*

$$\boldsymbol{\sigma}(t) = \mathcal{A}\boldsymbol{\varepsilon}(\dot{\boldsymbol{u}}(t)) + \mathcal{G}\boldsymbol{\varepsilon}(\boldsymbol{u}(t)) \qquad \forall t \in [0,T], \tag{3.24}$$

$$(\boldsymbol{\sigma}(t),\boldsymbol{\varepsilon}(\boldsymbol{v}))_Q + j(\dot{\boldsymbol{u}}(t),\boldsymbol{v}) = (\boldsymbol{f}(t),\boldsymbol{v})_V \qquad \forall\,\boldsymbol{v} \in V,\ t \in [0,T], \tag{3.25}$$

$$\boldsymbol{u}(0) = \boldsymbol{u}_0. \tag{3.26}$$

 The well-posedness of this problem is given by the following result.

Theorem 3.1. *Assume that (3.15)–(3.19) hold. Then, there exists $\beta_0 > 0$ which depends only on Ω, Γ_1, Γ_3 and $\mathcal{A}$ such that Problem PV has a unique solution $\{\boldsymbol{u}, \boldsymbol{\sigma}\}$ if*

$$|\beta|_{L^\infty(\Gamma_3)} < \beta_0. \tag{3.27}$$

Moreover, the solution satisfies

$$\boldsymbol{u} \in C^1([0,T]; V), \quad \boldsymbol{\sigma} \in C([0,T]; Q_1).$$

The proof of Theorem 3.1 is carried out in [2] in several steps, based on classical results for time-dependent nonlinear equations with strongly monotone operators followed by a fixed point argument. In the rest of the paper, we assume the conditions stated in Theorem 3.1 are satisfied so that contact problem PV has a unique solution.

We end this section with the remark that if v^* is large enough then $\beta = 1/(k_w v^*)$ is small enough and therefore condition (3.27) which guarantees the unique solvability of Problem PV is satisfied. We conclude that the mechanical problem (3.6)–(3.11) has a unique weak solution if the velocity of the foundation is large enough.

4 Fully discrete approximation

In this section we consider a fully discrete approximation of Problem PV. Let $V^h \subset V$ and $Q^h \subset Q$ denote finite dimensional spaces used to approximate the spaces V and Q satisfying $\boldsymbol{\varepsilon}(V^h) \subset Q^h$. Here $h > 0$ is a discretization parameter. Let $\mathcal{P}_{Q^h} : Q \to Q^h$ be the orthogonal projection operator defined through the relation

$$(\mathcal{P}_{Q^h}\boldsymbol{q}, \boldsymbol{q}^h)_Q = (\boldsymbol{q}, \boldsymbol{q}^h)_Q \qquad \forall\, \boldsymbol{q} \in Q,\ \forall \boldsymbol{q}^h \in Q^h.$$

The orthogonal projection operator is non-expansive, i.e.

$$|\mathcal{P}_{Q^h}\boldsymbol{q}|_Q \le |\boldsymbol{q}|_Q \qquad \forall\, \boldsymbol{q} \in Q.$$

This property will be used on various occasions.

In addition, we need a partition of the time interval $[0,T]$: $0 = t_0 < t_1 < \ldots < t_N = T$. We denote the step size $k_n = t_n - t_{n-1}$ for $n = 1, 2, \ldots, N$ and let $k = \max\limits_n k_n$ be the maximal step size. For a sequence $\{w_n\}_{n=0}^N$, we denote $\Delta w_n = w_n - w_{n-1}$ for the difference and $\delta w_n = \Delta w_n / k_n$ the corresponding divided difference. In this section no summation is considered over the repeated index n and c will denote positive constants which are independent of the parameters of discretization h and k.

The fully discrete approximation method is based on the backward Euler scheme. It has the following form.

Problem PV^{hk}**.** Find $\boldsymbol{u}^{hk} = \{\boldsymbol{u}_n^{hk}\}_{n=0}^N \subset V^h$ and $\boldsymbol{\sigma}^{hk} = \{\boldsymbol{\sigma}_n^{hk}\}_{n=1}^N \subset Q^h$ such that

$$\boldsymbol{\sigma}_n^{hh} = \mathcal{P}_{Q^h}\mathcal{A}\boldsymbol{\varepsilon}(\delta\boldsymbol{u}_n^{hk}) + \mathcal{P}_{Q^h}\mathcal{G}\boldsymbol{\varepsilon}(\boldsymbol{u}_n^{hk}), \tag{4.1}$$

$$(\boldsymbol{\sigma}_n^{hk}, \boldsymbol{\varepsilon}(\boldsymbol{w}^h))_Q + j(\delta\boldsymbol{u}_n^{hk}, \boldsymbol{w}^h) = (\boldsymbol{f}_n, \boldsymbol{w}^h)_V \quad \forall\, \boldsymbol{w}^h \in V^h, \tag{4.2}$$

$$\boldsymbol{u}_0^{hk} = \boldsymbol{u}_0^h. \tag{4.3}$$

Using similar arguments to those employed in [2], we have the following.

Theorem 4.1. *Assume that (3.15)–(3.19) and (3.27) hold. Then Problem* PV^{hk} *has a unique solution.*

Remark 4.2. In practice, the fixed point algorithm used to show the existence and uniqueness of solution for Problem PV^{hk} is directly applied. To solve the semilinear equality obtained, a penalty-duality algorithm introduced in [4] is suggested. Details are given in Section 5.

We now derive error estimates for the fully discrete solution. Let $n \in \{1, 2, ..., N\}$ and let $\boldsymbol{\sigma}_n = \boldsymbol{\sigma}(t_n)$, $\boldsymbol{v}_n = \dot{\boldsymbol{u}}(t_n)$, $\boldsymbol{u}_n = \boldsymbol{u}(t_n)$, $\boldsymbol{v}_n^{hk} = \delta\boldsymbol{u}_n^{hk}$. By integration of (3.24) we get

$$\boldsymbol{\sigma}(t) = \mathcal{A}\boldsymbol{\varepsilon}(\boldsymbol{v}(t)) + \mathcal{G}\boldsymbol{\varepsilon}(\int_0^t \boldsymbol{v}(s)ds + \boldsymbol{u}_0) \qquad \forall t \in [0, T], \tag{4.4}$$

where $\boldsymbol{v}(t) = \dot{\boldsymbol{u}}(t)$. Substracting now (4.4) at $t = t_n$ to (4.1), we obtain

$$\begin{aligned}\boldsymbol{\sigma}_n - \boldsymbol{\sigma}_n^{hk} &= (I - \mathcal{P}_{Q^h})\mathcal{A}\boldsymbol{\varepsilon}(\boldsymbol{v}_n) + \mathcal{P}_{Q^h}(\mathcal{A}\boldsymbol{\varepsilon}(\boldsymbol{v}_n) - \mathcal{A}\boldsymbol{\varepsilon}(\boldsymbol{v}_n^{hk})) \\ &+(I - \mathcal{P}_{Q^h})\mathcal{G}\boldsymbol{\varepsilon}(\boldsymbol{u}_n) + \mathcal{P}_{Q^h}[\mathcal{G}\boldsymbol{\varepsilon}(\int_0^{t_n} \boldsymbol{v}(s)ds + \boldsymbol{u}_0) - \mathcal{G}\boldsymbol{\varepsilon}(\sum_{j=1}^n k_j\boldsymbol{v}_j^{hk} + \boldsymbol{u}_0^h)],\end{aligned}$$

where I is the identity operator on Q.

Thus,

$$\begin{aligned}|\boldsymbol{\sigma}_n - \boldsymbol{\sigma}_n^{hk}|_Q \le |(I - \mathcal{P}_{Q^h})\mathcal{A}\boldsymbol{\varepsilon}(\boldsymbol{v}_n)|_Q + c|\boldsymbol{v}_n - \boldsymbol{v}_n^{hk}|_V + |\boldsymbol{u}_0 - \boldsymbol{u}_0^h|_V \\ +|(I - \mathcal{P}_{Q^h})\mathcal{G}\boldsymbol{\varepsilon}(\boldsymbol{u}_n)|_Q + |\int_0^{t_n} \boldsymbol{v}(s)ds - \sum_{j=1}^n k_j\boldsymbol{v}_j^{hk}|_V.\end{aligned} \tag{4.5}$$

Plugging (4.4) into (3.25) at time $t = t_n$, (4.1) into (4.2) and substracting the obtained

equalities, we deduce that

$$(\mathcal{A}\boldsymbol{\varepsilon}(\boldsymbol{v}_n) - \mathcal{A}\boldsymbol{\varepsilon}(\boldsymbol{v}_n^{hk}), \boldsymbol{\varepsilon}(\boldsymbol{w}^h))_Q + j(\boldsymbol{v}_n, \boldsymbol{w}^h) - j(\boldsymbol{v}_n^{hk}, \boldsymbol{w}^h) =$$
$$-(\mathcal{G}\boldsymbol{\varepsilon}(\int_0^{t_n} \boldsymbol{v}(s)ds + \boldsymbol{u}_0) - \mathcal{G}\boldsymbol{\varepsilon}(\sum_{j=1}^{n} k_j \boldsymbol{v}_j^{hk} + \boldsymbol{u}_0^h), \boldsymbol{\varepsilon}(\boldsymbol{w}^h))_Q \quad \forall \boldsymbol{w}^h \in V^h.$$

Therefore

$$\begin{aligned}
&(\mathcal{A}\boldsymbol{\varepsilon}(\boldsymbol{v}_n) - \mathcal{A}\boldsymbol{\varepsilon}(\boldsymbol{v}_n^{hk}), \boldsymbol{\varepsilon}(\boldsymbol{v}_n) - \boldsymbol{\varepsilon}(\boldsymbol{v}_n^{hk}))_Q + j(\boldsymbol{v}_n, \boldsymbol{v}_n - \boldsymbol{v}_n^{hk}) - j(\boldsymbol{v}_n^{hk}, \boldsymbol{v}_n - \boldsymbol{v}_n^{hk}) \\
&= (\mathcal{A}\boldsymbol{\varepsilon}(\boldsymbol{v}_n) - \mathcal{A}\boldsymbol{\varepsilon}(\boldsymbol{v}_n^{hk}), \boldsymbol{\varepsilon}(\boldsymbol{v}_n) - \boldsymbol{\varepsilon}(\boldsymbol{w}^h))_Q + j(\boldsymbol{v}_n, \boldsymbol{v}_n - \boldsymbol{w}^h) - j(\boldsymbol{v}_n^{hk}, \boldsymbol{v} - \boldsymbol{w}^h) \\
&+(\mathcal{G}\boldsymbol{\varepsilon}(\int_0^{t_n} \boldsymbol{v}(s)ds + \boldsymbol{u}_0) - \mathcal{G}\boldsymbol{\varepsilon}(\sum_{j=1}^{n} k_j \boldsymbol{v}_j^{hk} + \boldsymbol{u}_0^h), \boldsymbol{\varepsilon}(\boldsymbol{v}_n - \boldsymbol{w}^h))_Q \\
&-(\mathcal{G}\boldsymbol{\varepsilon}(\int_0^{t_n} \boldsymbol{v}(s)ds + \boldsymbol{u}_0) - \mathcal{G}\boldsymbol{\varepsilon}(\sum_{j=1}^{n} k_j \boldsymbol{v}_j^{hk} + \boldsymbol{u}_0^h), \boldsymbol{\varepsilon}(\boldsymbol{v}_n - \boldsymbol{v}_n^{hk}))_Q \qquad \forall \boldsymbol{w}^h \in V^h.
\end{aligned}$$

It can be verified that

$$\begin{aligned}
(\mathcal{A}\boldsymbol{\varepsilon}(\boldsymbol{v}_n) - \mathcal{A}\boldsymbol{\varepsilon}(\boldsymbol{v}_n^{hk}), \boldsymbol{\varepsilon}(\boldsymbol{v}_n) - \boldsymbol{\varepsilon}(\boldsymbol{v}_n^{hk}))_Q \quad &+j(\boldsymbol{v}_n, \boldsymbol{v}_n - \boldsymbol{v}_n^{hk}) - j(\boldsymbol{v}_n^{hk}, \boldsymbol{v}_n - \boldsymbol{v}_n^{hk}) \\
&\geq (m_{\mathcal{A}} - C_0^2\beta_0)|\boldsymbol{v}_n - \boldsymbol{v}_n^{hk}|_V^2,
\end{aligned}$$

where β_0 is chosen in such a way that $m_{\mathcal{A}} - C_0^2\beta_0 > 0$, and C_0 is given by (3.14).

After some algebra, we then obtain the inequality

$$|\boldsymbol{v}_n - \boldsymbol{v}_n^{hk}|_V \leq c\Big(|\boldsymbol{v}_n - \boldsymbol{w}^h|_V + I_n + \sum_{j=1}^{n} k_j e_j\Big) \quad \forall \boldsymbol{w}^h \in V^h, \tag{4.6}$$

where

$$I_n = |\int_0^{t_n} \boldsymbol{v}(s)ds - \sum_{j=1}^{n} k_j \boldsymbol{v}_j|_V, \quad e_j = |\boldsymbol{v}_j - \boldsymbol{v}_j^{hk}|_V.$$

The following lemma was proved in [5].

Lemma 4.3. *Assume $\{g_n\}_{n=1}^N$ and $\{e_n\}_{n=1}^N$ are two sequences of non-negative numbers satisfying*

$$e_n \leq cg_n + c\sum_{j=1}^{n} k_j e_j,$$

then

$$\max_{1\leq n\leq N} e_n \leq \max_{1\leq n\leq N} g_n.$$

Moreover, for the quadrature error I_n we have

$$I_n \leq \sum_{j=1}^{n} \int_{t_{j-1}}^{t_j} |\boldsymbol{v}(s) - \boldsymbol{v}_j|_V ds \leq ct_n k|\dot{\boldsymbol{v}}|_{L^\infty(0,T;V)} \tag{4.7}$$

if $\dot{\boldsymbol{v}} \in L^\infty(0,T;V)$. Applying Lemma 4.3 to (4.6) we obtain the following error estimate:

$$\max_{1\le n\le N} |\boldsymbol{v}_n - \boldsymbol{v}_n^{hk}|_V \le c\Big(k|\dot{\boldsymbol{v}}|_{L^\infty(0,T;V)} + \max_{1\le n\le N} \inf_{\boldsymbol{w}^h\in V^h} |\boldsymbol{v}_n - \boldsymbol{w}^h|_V\Big).$$

Using now the definitions of $\boldsymbol{v}_n$ and $\boldsymbol{v}_n^{hk}$ we find

$$|\boldsymbol{u}_n - \boldsymbol{u}_n^{hk}|_V \le c(I_n + \max_{1\le j\le N} |\boldsymbol{v}_j - \boldsymbol{v}_h^{hk}|_V + |\boldsymbol{u}_0 - \boldsymbol{u}_0^h|_V). \tag{4.8}$$

Therefore, from (4.5), (4.7) and (4.8) we obtain the following result.

Theorem 4.4. *Let $\{\boldsymbol{u},\boldsymbol{\sigma}\} \in C^1([0,T];V)\times C([0,T];Q_1)$ be the solution of Problem PV and $\{(\boldsymbol{u}_n^{hk},\boldsymbol{\sigma}_n^{hk})\}_{n=0}^N \subset V^h\times Q^h$ be the solution of fully discrete problem PV^{hk}. Assume conditions (3.15)–(3.19) and (3.27) hold. Let us suppose $\ddot{\boldsymbol{u}} \in L^\infty(0,T;V)$. Then we have the following error estimate:*

$$\begin{aligned}\max_{1\le n\le N} \ & \{|\boldsymbol{u}_n - \boldsymbol{u}_n^{hk}|_V + |\boldsymbol{\sigma}_n - \boldsymbol{\sigma}_n^{hk}|_Q\} \le c\Big(|\boldsymbol{u}_0 - \boldsymbol{u}_0^h|_V \\ & + k|\ddot{\boldsymbol{u}}|_{L^\infty(0,T;V)} + \max_{1\le n\le N} \inf_{\boldsymbol{w}^h\in V^h} |\dot{\boldsymbol{u}}_n - \boldsymbol{w}^h|_V \\ & + \max_{1\le n\le N}\{|(I-\mathcal{P}_{Q^h})\mathcal{A}\boldsymbol{\varepsilon}(\dot{\boldsymbol{u}}_n)|_Q + |(I-\mathcal{P}_{Q^h})\mathcal{G}\boldsymbol{\varepsilon}(\boldsymbol{u}_n)|_Q\}\Big).\end{aligned}$$

As a consequence of Theorem 4.4 we obtain the following result.

Corollary 4.5. *Assume that conditions (3.15)–(3.19) and (3.27) hold. Then under the additional assumptions*

$$\begin{aligned}&|\boldsymbol{u}_0 - \boldsymbol{u}_0^h|_V \le ch, \quad \inf_{\boldsymbol{w}^h\in V^h} |\boldsymbol{w} - \boldsymbol{w}^h|_V \le ch \quad \forall \boldsymbol{w} \in V\cap[H^2(\Omega)]^d,\\ &|(I-\mathcal{P}_{Q^h})(\boldsymbol{\tau})|_Q \le ch \quad \forall \boldsymbol{\tau}\in Q, \quad \dot{\boldsymbol{u}} \in L^\infty(0,T;[H^2(\Omega)]^d),\end{aligned}$$

we obtain the following error estimate:

$$\max_{1\le n\le N}\{|\boldsymbol{u}_n - \boldsymbol{u}_n^{hk}|_V + |\boldsymbol{\sigma}_n - \boldsymbol{\sigma}_n^{hk}|_Q\} \le c(h+k).$$

5 Numerical results

In order to show the perfomance of the numerical method described in the above section, some numerical experiments have been done in the study of one and two dimensional test problems. In this section, we describe the complete algorithm to solve Problem PV^{hk} and we resume some numerical results which exhibit the perfomance of the algorithm.

5.1 Algorithm to solve Problem PV^{hk}

Sustituting (4.1) into (4.2), we obtain for $n = 1, 2, \ldots, N$:

$$(P_{Q^h}[\mathcal{A}\boldsymbol{\varepsilon}(\delta \boldsymbol{u}_n^{hk}) + \mathcal{G}\boldsymbol{\varepsilon}(\boldsymbol{u}_n^{hk})], \boldsymbol{\varepsilon}(\boldsymbol{w}^h))_Q + j(\delta \boldsymbol{u}_n^{hk}, \boldsymbol{w}^h) = (\boldsymbol{f}_n, \boldsymbol{w}^h)_V \quad \forall\, \boldsymbol{w}^h \in V^h, \quad (5.1)$$

where $\boldsymbol{u}_0^{hk} = \boldsymbol{u}_0^h$ is given.

Using the properties of the orthogonal operator P_{Q^h} defined in Section 4 and defining again $\boldsymbol{v}_n^{hk} = \delta \boldsymbol{u}_n^{hk}$, we obtain that (5.1) is equivalent to:

$$(\mathcal{A}\boldsymbol{\varepsilon}(\boldsymbol{v}_n^{hk}) + \mathcal{G}\boldsymbol{\varepsilon}(\boldsymbol{u}_{n-1}^{hk} + k_n \boldsymbol{v}_n^{hk}), \boldsymbol{\varepsilon}(\boldsymbol{w}^h))_Q + j(\boldsymbol{v}_n^{hk}, \boldsymbol{w}^h) = (\boldsymbol{f}_n, \boldsymbol{w}^h)_V \ \ \forall \boldsymbol{w}^h \in V^h. \quad (5.2)$$

Then, the element $\boldsymbol{u}_n^{hk}$ is obtained by

$$\boldsymbol{u}_n^{hk} = \boldsymbol{u}_{n-1}^{hk} + k_n \boldsymbol{v}_n^{hk}, \quad n = 1, 2, \ldots, N.$$

In order to solve problem (5.2), the first step is to apply the fixed point argument introduced in [2] to prove the existence and uniqueness of solution.

For $\boldsymbol{\eta}_0^{hk} \in Q^h$ given, let $\{(\boldsymbol{v}_{\boldsymbol{\eta}_r^{hk}}, \boldsymbol{\eta}_r^{hk})\}_{r\geq 0} \subset V^h \times Q^h$ be the sequence obtained solving problems (5.3)-(5.4) described below:

$$\left.\begin{array}{l} \boldsymbol{v}_{\boldsymbol{\eta}_r^{hk}} \in V^h, \\ (\mathcal{A}\boldsymbol{\varepsilon}(\boldsymbol{v}_{\boldsymbol{\eta}_r^{hk}}) + \boldsymbol{\eta}_r^{hk}, \boldsymbol{\varepsilon}(\boldsymbol{w}^h))_Q + j(\boldsymbol{v}_{\boldsymbol{\eta}_r^{hk}}, \boldsymbol{w}^h) = (\boldsymbol{f}_n, \boldsymbol{w}^h)_V \quad \forall\, \boldsymbol{w}^h \in V^h, \end{array}\right\} \quad (5.3)$$

$$\boldsymbol{\eta}_{r+1}^{hk} = \mathcal{G}\boldsymbol{\varepsilon}(\boldsymbol{u}_{n-1}^{hk} + k_n \boldsymbol{v}_{\boldsymbol{\eta}_r^{hk}}). \quad (5.4)$$

We have (see [2]),

$$\lim_{r\to\infty} \boldsymbol{v}_{\boldsymbol{\eta}_r^{hk}} = \boldsymbol{v}_n^{hk}.$$

To solve the semilinear equality (5.3) we used a penalty-duality algorithm introduced in [4]. In order to describe it, we remark that

$$j(\boldsymbol{u}, \boldsymbol{v}) = \int_{\Gamma_3} \beta |u_\nu| v_\nu da = \int_{\Gamma_3} \beta (u_\nu^+ - u_\nu^-) v_\nu \, da = \int_{\Gamma_3} \beta u_\nu^+ v_\nu da - \int_{\Gamma_3} \beta u_\nu^- v_\nu da,$$

where $u_\nu^+ = \max\{u_\nu, 0\}$ and $u_\nu^- = \min\{u_\nu, 0\}$.

Let $E^h \subset L^2(\Gamma_3)$ and $K^h \subset E^h$ be defined by

$$E^h = \{q^h \in L^2(\Gamma_3)\,;\, q^h_{|C} \in P_1(C) \quad \forall C \in \theta^h\},$$
$$K^h = \{q^h \in E^h\,;\, q^h_{|C} \leq 0 \text{ in } C \quad \forall C \in \theta^h\},$$

where θ^h is the induced triangulation of $\mathcal{T}^h$ on Γ_3.

Let $B : V^h \to E^h$ be the operator

$$B\boldsymbol{v}^h = v_\nu^h \quad \forall \boldsymbol{v}^h \in V^h,$$

and let $B^* : E^h \to V^h$ be the operator given by

$$(B^* q^h, \boldsymbol{v}^h)_V = \int_{\Gamma_3} \beta q^h v_\nu^h da \quad \forall \boldsymbol{v}^h \in V^h, \forall q^h \in E^h.$$

It is well known that problem (5.3) is equivalent to minimize on V^h the functional $J : V^h \to \mathbb{R}$ defined by

$$J(\boldsymbol{v}^h) = f(\boldsymbol{v}^h) - g(B\boldsymbol{v}^h),$$

where $f : V^h \to (-\infty, \infty]$ and $g : E^h \to (-\infty, \infty]$ have the following form,

$$f(\boldsymbol{v}^h) = \frac{1}{2}(\mathcal{A}\boldsymbol{\varepsilon}(\boldsymbol{v}^h), \boldsymbol{\varepsilon}(\boldsymbol{v}^h))_Q + \frac{1}{2}\int_{\Gamma_3} \beta[(v_\nu^h)^+]^2 da - (\boldsymbol{f}_n, \boldsymbol{v}^h)_V - (\boldsymbol{\eta}_r^{hk}, \boldsymbol{\varepsilon}(\boldsymbol{v}^h))_Q,$$
$$g(q^h) = \frac{1}{2}\int_{\Gamma_3} \beta[(q^h)^-]^2 da.$$

According to [4], the following iterative algorithm is considered:

Given $p_0^{hk} \in E^h$ and $\lambda > 0$, let $\{(\boldsymbol{v}_m^{hk}, p_m^{hk})\}_{m\geq 0} \subset V^h \times E^h$ be the sequence obtained solving the following problems:

$$B^* p_m^{hk} \in \partial f(\boldsymbol{v}_{m+1}^{hk}), \tag{5.5}$$

$$p_{m+1}^{hk} = g'_\lambda(B\boldsymbol{v}_{m+1}^{hk} + \lambda p_m^{hk}), \tag{5.6}$$

where ∂f denotes the subdifferential operator of f and g'_λ is the Yosida's approximation of the maximal monotone operator ∂g. Then (see [4]):

$$\lim_{m\to\infty} \boldsymbol{v}_m^{hk} = \boldsymbol{v}_{\boldsymbol{\eta}_r^{hk}}.$$

We can see that problem (5.5)-(5.6) is equivalent to the following.

$$\left.\begin{aligned} &\boldsymbol{v}_{m+1}^{hk} \in V^h, \\ &(\mathcal{A}\boldsymbol{\varepsilon}(\boldsymbol{v}_{m+1}^{hk}), \boldsymbol{\varepsilon}(\boldsymbol{v}^h - \boldsymbol{v}_{m+1}^{hk}))_Q + \frac{1}{2}\int_{\Gamma_3} \beta[(v_\nu^h)^+]^2 da - \frac{1}{2}\int_{\Gamma_3} \beta[(\boldsymbol{v}_{m+1}^{hk})_\nu^+]^2 da \\ &\qquad \geq (\boldsymbol{f}_n, \boldsymbol{v}^h - \boldsymbol{v}_{m+1}^{hk})_V - (\boldsymbol{\eta}_r^{hk}, \boldsymbol{\varepsilon}(\boldsymbol{v}^h - \boldsymbol{v}_{m+1}^{hk}))_Q \\ &\qquad\qquad + \int_{\Gamma_3} \beta p_m^{hk}(v_\nu^h - (\boldsymbol{v}_{m+1}^{hk})_\nu)\, da \quad \forall \boldsymbol{v}^h \in V^h, \end{aligned}\right\} \tag{5.7}$$

$$p_{m+1}^{hk} = \frac{1}{\lambda}(B\boldsymbol{v}_{m+1}^{hk} + \lambda p_m^{hk})^-. \tag{5.8}$$

Finally, problem (5.7) is similar to the normal compliance contact problem in elasticity (see [13]). To solve it, we used the following penalty-duality algorithm previously introduced in [14]:

Given $q_0^{hk} \in E^h$ *and* $\omega \in \mathbb{R}$, $\omega > 0$, *let* $\{(\boldsymbol{v}_l^{hk}, q_l^{hk})\}_{l\geq 0} \subset V^h \times E^h$ *be the sequence obtained solving the following problems:*

$$\left.\begin{aligned} &\boldsymbol{v}_l^{hk} \in V^h, \\ &(\mathcal{A}\boldsymbol{\varepsilon}(\boldsymbol{v}_l^{hk}), \boldsymbol{\varepsilon}(\boldsymbol{w}^h))_Q \quad +\omega \int_{\Gamma_3} \beta(\boldsymbol{v}_l^{hk})_\nu w_\nu^h da = (\boldsymbol{f}_n, \boldsymbol{w}^h)_V \\ &\quad -(\boldsymbol{\eta}_r^{hk}, \boldsymbol{\varepsilon}(\boldsymbol{w}^h))_Q + \int_{\Gamma_3} \beta p_m^{hk} w_\nu^h \, da - \int_{\Gamma_3} \beta \, q_l^{hk} w_\nu^h da \quad \forall \, \boldsymbol{w}^h \in V^h, \end{aligned}\right\} \tag{5.9}$$

$$\left.\begin{aligned} &\tilde{q}_l^{hk} = 2(\boldsymbol{v}_l^{hk})_\nu + \frac{1}{\omega} q_l^{hk}, \\ &q_{l+1}^{hk} = \frac{\omega}{1+\omega}[(1-\omega)\tilde{q}_l^{hk} - 2P_{K^h}(\tilde{q}_l^{hk})], \end{aligned}\right\} \tag{5.10}$$

where P_{K^h} is the orthogonal $L^2(\Gamma_3)$-projection operator over the convex set K^h. We obtain (see [14]):

$$\lim_{l\to\infty} \boldsymbol{v}_l^{hk} = \boldsymbol{v}_{m+1}^{hk}.$$

5.2 A one-dimensional test-problem

We consider a viscoelastic rod $\Omega = (0, L)$ which is fixed at its left end $x = 0$ and is subjected to the action of a body force of density $f_0(x, t)$ in the x-direction (see Figure 1). Its right end $x = L$ is in contact with a rigid moving obstacle which produces the

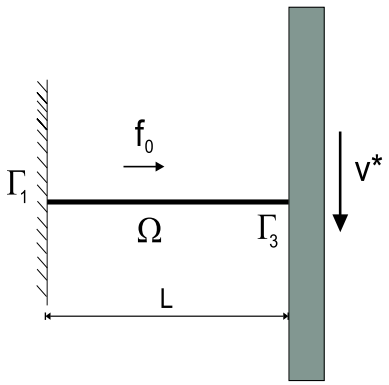


Figure 1: Test 1: Sliding contact of a viscoelastic rod.

wear of the rod. This problem corresponds to Problem P with $\Omega = (0, L)$, $\Gamma_1 = \{0\}$, $\Gamma_2 = \emptyset$, $\Gamma_3 = \{L\}$. We use a linear viscoelastic constitutive law, i.e.

$$\sigma = a\,\varepsilon(\dot{u}) + g\,\varepsilon(u).$$

Here $\varepsilon(u) = \dfrac{\partial u}{\partial x}$, while a and g are material constants independants on x and t, $a > 0$.

A complete description of this problem is the following one.

Problem $T1D$. *Find a displacement field* $u : [0, L] \times [0, T] \to \mathbb{R}$ *and a stress field* $\sigma : [0, L] \times [0, T] \to \mathbb{R}$ *such that:*

$$\sigma(x,t) = a\frac{\partial^2 u(x,t)}{\partial x \partial t} + g\frac{\partial u(x,t)}{\partial x} \qquad \text{in } (0,L)\times(0,T), \tag{5.11}$$

$$-\frac{\partial \sigma(x,t)}{\partial x} = f_0(x,t) \qquad \text{in } (0,L)\times(0,T), \tag{5.12}$$

$$u(0,t) = 0 \qquad \text{for } \ t \in (0,T), \tag{5.13}$$

$$\sigma(1,t) = -\beta\Big|\frac{\partial u(1,t)}{\partial t}\Big| \qquad \text{for } t \in (0,T), \tag{5.14}$$

$$u(x,0) = u_0(x) \qquad \text{in } \ (0,L). \tag{5.15}$$

For computation we have used the following data:

$$L = 1\,m, \quad T = 10s., \quad a = 100\,N\cdot s/m, \quad g = 1\,N/m,$$
$$f_0(x,t) = 99Ae^{-t} \quad \forall x \in (0,1),\ t \in [0,10],$$
$$\beta = 10^{-4}\,N\cdot s/m, \quad u_0(x) = A\tfrac{x^2}{2} + x \quad \forall x \in (0,1),$$
$$A = -\frac{9899.99}{9899.995}.$$

The exact solution of the above problem is

$$u(x,t) = \Big(A\frac{x^2}{2} + x\Big)e^{-t}, \tag{5.16}$$

$$\sigma(x,t) = -99\Big(Ax + 1\Big)e^{-t}. \tag{5.17}$$

We used a discretization by continuous piecewise affine functions for the space V^h and by piecewise constant functions for the space Q^h. Since $\boldsymbol{\varepsilon}(V^h) \subset Q^h$, in (4.1) no projection is needed.

In Table 1 we provide the exact error values for several discretization parameters h and k. We note that the asymptotic behavior

$$\max_{1\le n\le N}(|\boldsymbol{u}_n - \boldsymbol{u}_n^{hk}|_V + |\boldsymbol{\sigma}_n - \boldsymbol{\sigma}_n^{hk}|_Q) \le c(h+k),$$

$k \downarrow \vert h \rightarrow$	0.1	0.05	0.025	0.01	0.005	0.0025	0.001
0.1	0.026276	0.013411	0.007236	0.004078	0.003392	0.003198	0.003142
0.05	0.027474	0.013749	0.006910	0.002862	0.001593	0.001061	0.000855
0.025	0.028167	0.014079	0.007041	0.002822	0.001422	0.000703	0.000349
0.01	0.028594	0.014291	0.007145	0.002858	0.014210	0.000715	0.000287
0.005	0.028738	0.014363	0.007181	0.002872	0.001436	0.000718	0.000287
0.0025	0.028809	0.014399	0.007199	0.002879	0.001439	0.000719	0.000287
0.001	0.028853	0.014421	0.007209	0.002883	0.001442	0.000721	0.000288

Table 1: Problem $T1D$: Exact error values.

derived in Corollary 4.5, is obtained with the constant $c \approx 0.13874157$, independent of h and k (see Figure 2 (left-hand side)). The evolution of the wear, obtained from (3.3), is shown on the right-hand side of Figure 2. For computations we used here the values $k = h = 0.01$.

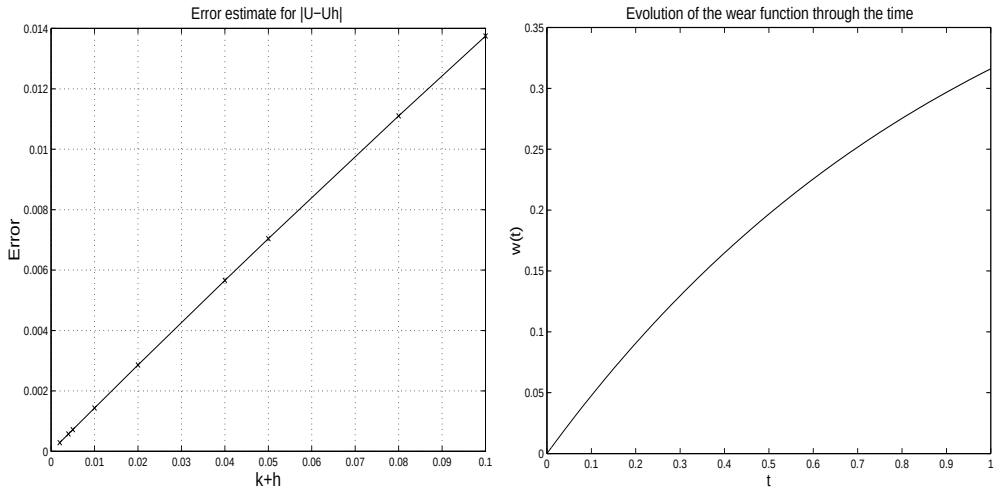


Figure 2: Problem $T1D$: Asymptotic constant of error and evolution of the wear function through the time.

5.3 A two-dimensional test problem

As a two-dimensional example of Problem P, we consider the plane stress viscoelasticity problem for an isotropic material shown in Figure 3. The physical setting corresponds to a three-dimensional linear viscoelastic body of cross-section $\Omega = (0,6)\times(0,6)$ submitted to the action of vertical forces in such a way that a plane stress hypothesis is assumed. A linearly decreasing traction force is supposed to act on the part $\Gamma_2 = [0,6] \times \{6\}$ of the boundary and no body forces act in Ω. The horizontal displacements on the lateral surface are supposed to vanish, i.e. $\Gamma_1 = \{0,6\}\times[0,6]$. Finally, the body is in sliding contact with a rigid moving foundation on $\Gamma_3 = [0,6]\times\{0\}$.

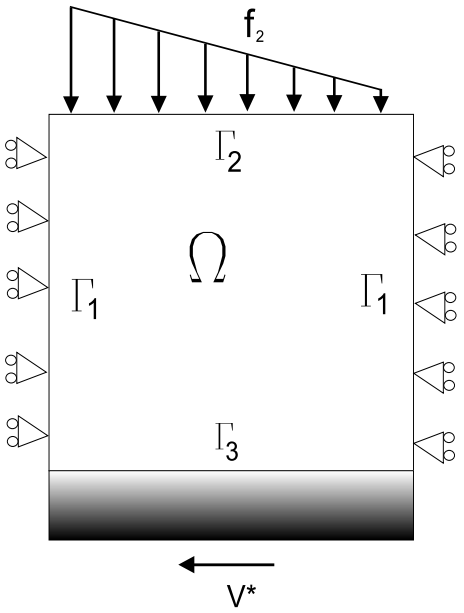


Figure 3: Test 2: Contact of a 2D viscoelastic body.

A permanent contact with the foundation is maintained and the wear is then produced. We assume the wear coefficient to be constant on Γ_3 and therefore the data β is constant on Γ_3 too.

The elasticity tensor $\mathcal{G}$ is then defined by:

$$(\mathcal{G}\boldsymbol{\tau})_{\alpha\beta} = \frac{E\kappa}{1-\kappa^2}(\tau_{11}+\tau_{22})\delta_{\alpha\beta} + \frac{E}{1+\kappa}\tau_{\alpha\beta}, \qquad 1 \le \alpha, \beta \le 2,$$

where E is the Young's modulus, κ the Poisson's ratio of the material and $\delta_{\alpha\beta}$ denotes the Kronecker symbol. The viscosity tensor $\mathcal{A}$ has a similar form, i.e.

$$(\mathcal{A}\boldsymbol{\tau})_{\alpha\beta} = \mu(\tau_{11}+\tau_{22})\delta_{\alpha\beta} + \eta\tau_{\alpha\beta}, \qquad 1 \le \alpha, \beta \le 2,$$

where μ and η are viscosity constants. Recall also that the von Mises norm for a plane stress field $\boldsymbol{\tau} = (\tau_{\alpha\beta})$ is given by

$$\|\boldsymbol{\tau}\| = \left(\tau_{11}^2 + \tau_{22}^2 - \tau_{11}\tau_{22} + 3\tau_{12}^2\right)^{\frac{1}{2}}.$$

For computation we have used the following data:

$$T = 1\,\mathrm{s}, \quad \boldsymbol{f}_0 = \mathbf{0}\,N/m^3, \quad \boldsymbol{f}_2(x_1, x_2, t) = (0, -10(6-x_1)t)\,N/m^2,$$
$$\boldsymbol{\sigma}_0 = \mathbf{0}\ N/m^2, \quad \boldsymbol{u}_0 = \mathbf{0}\,m, \quad \beta = 10^{-2}\,N\cdot s/m^2,$$
$$E = 100N/m^2, \quad \kappa = 0.3, \quad \mu = 32.967N\cdot s/m^2, \quad \eta = 23.077N\cdot s/m^2.$$

72 As in the one dimensional test problem, we used continuous piecewise affine functions to approximate the space V and piecewise constant functions to approximate the space Q. Again, $\boldsymbol{\varepsilon}(V^h) \subset Q^h$ and therefore no projections are needed in the discrete constitutive law (4.1). Moreover, for the time discretization, we used a uniform time step with $k = 0.01$.

The amplified deformed mesh, the von Mises norm of stress at final time $T = 1\,s$ and the initial boundary are shown in Figure 4. We used again (3.3) to obtain the wear function. In Figure 5 we show the wear of the contact boundary for several time values as well as the evolution of the wear of the central contact node $\boldsymbol{x} = (3, 0)$.

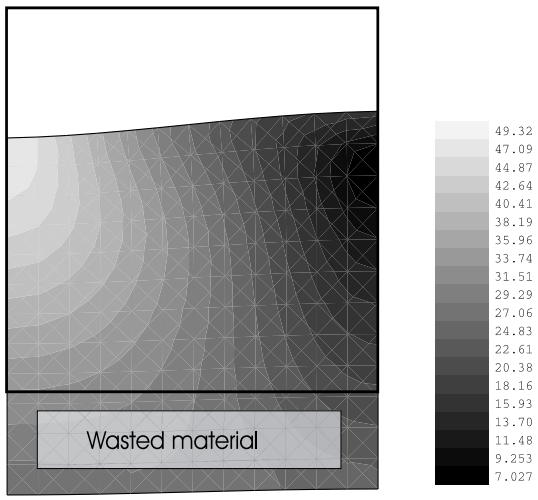


Figure 4: Amplified (by 60) deformed mesh and von Mises stress norm at final time $T = 1\,s$ and initial boundary.

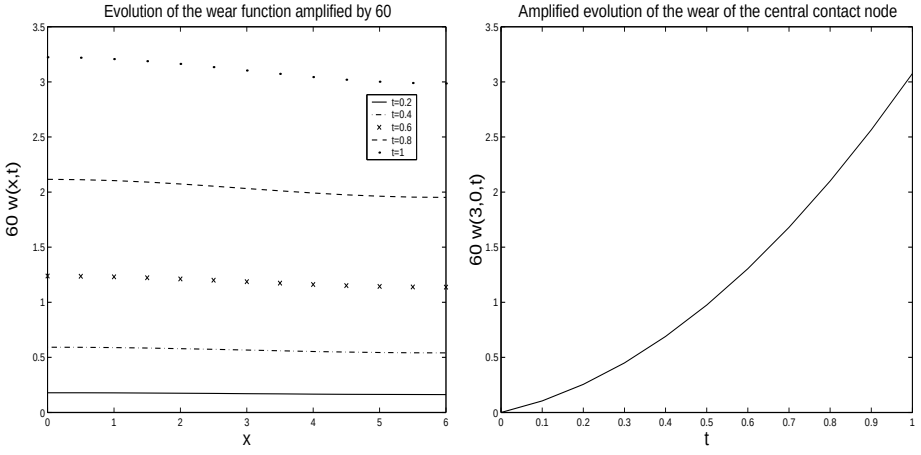


Figure 5: Wear of the contact boundary for several time values amplified by 60 and evolution of the wear of the central contact node.

Acknowledgements

The work of the first and third authors was partially supported by DGESIC-Spain (Project BFM2003-05357).

References

[1] M. Burguera and J. M. Viaño, Numerical solving of frictionless contact problems in perfect plastic bodies, *Computer Methods in Applied Mechanics and Engineering*, **120** (1995), 303–322.

[2] C. Ciulcu, T.H. Hoarau-Mantel and M. Sofonea, Viscoelastic sliding frictional contact problems with wear. Lyapunov's methods in stability and control. *Mathematical and Computer Modelling,* **36**(7-8) (2002), 861–874.

[3] G. Duvaut and J.L. Lions, *Inequalities in Mechanics and Physics*, Springer-Verlag, Berlin, 1976.

[4] E. Fernández-Cara and C. Moreno, Critical point approximation through exact regularization. *Mathematics of Computation*, **50** (1988), 139–153.

[5] W. Han and M. Sofonea, Numerical analysis of a frictionless contact problem for elastic-viscoplastic materials, *Computer Methods in Applied Mechanics and Engineering*, **190** (2000), 179–191.

[6] I. R. Ionescu, M. Sofonea, *Functional and Numerical Methods in Viscoplasticity*, Oxford University Press, Oxford, 1993.

[7] J. Nečas and I. Hlavaček, *Mathematical Theory of Elastic and Elastoplastic Bodies: An Introduction*. Elsevier, Amsterdam, 1981.

[8] P. D. Panagiotopoulos, *Inequality Problems in Mechanical and Applications*. Birkhäuser, Basel 1985.

[9] M. Rochdi, M. Shillor and M. Sofonea, Quasistatic nonlinear viscoelastic contact with normal compliance and friction, *Journal of Elasticity*, **51** (1998), 105-126.

[10] M. Rochdi, M. Shillor and M. Sofonea, A quasistatic contact problem with directional friction and damped response. *Applicable Analysis*, **68** (1998), 409–422.

[11] N. Strömberg, *Continuum Thermodynamics of Contact, Friction and Wear*, Ph.D. Thesis, Linköping University, Sweden, 1995.

[12] N. Strömberg, L. Johansson and A. Klarbring, Derivation and analysis of a generalized standard model for contact friction and wear. *International Journal of Solids Structures*, **33** (1996), 1817–1836.

[13] J.M. Viaño, Análisis de un método numérico con elementos finitos para problemas de contacto unilateral sin rozamiento en elasticidad. Parte I: Formulación física y matemática de los problemas. *Revista Internacional de Métodos Numéricos para Cálculo y Diseño en Ingeniería*, **1** (1986), 79-93.

[14] J.M. Viaño, Análisis de un método numérico con elementos finitos para problemas de contacto unilateral sin rozamiento en elasticidad. Parte II: Aproximación y resolución de los problemas discretos. *Revista Internacional de Métodos Numéricos para Cálculo y Diseño en Ingeniería*, **2** (1986), 63-86.

JOURNAL OF CONCRETE AND APPLICABLE MATHEMATICS,VOL.3,NO.1,75-89,2005,COPYRIGHT 2005 EUDOXUS PRESS,LLC

Characterization of A Class of Band-limited Wavelets

Biswaranjan Behera
Department of Mathematics
Indian Institute of Technology Kanpur
208016, India.
Current address:
Department of Mathematics
Indian Institute of Technology Delhi
Hauz Khas, New Delhi
110016 India.
Email: biswa@maths.iitd.ac.in

Shobha Madan
Department of Mathematics
Indian Institute of Technology Kanpur
208016, India.
Email: madan@iitk.ac.in

Key words: Wavelet; MRA-wavelet; multiresolution analysis; dimension function.
2000 AMS Subject Classification: 42C40.

Abstract

For each integer $n \geq 2$, we construct a subset S_n of $\mathbb{R}$ and characterize all wavelets of $L^2(\mathbb{R})$ whose Fourier transform is supported in S_n. This result generalizes the characterization of a class of wavelets given by E. Hernández and G. Weiss. Wavelets are usually constructed by using a multiresolution analysis (MRA); but there are wavelets which cannot be constructed through an MRA. We show that the wavelets associated with S_n are non-MRA wavelets, if $n \geq 3$. The dimension functions of these wavelets are also computed explicitly.

1 Introduction

A function $\psi \in L^2(\mathbb{R})$ is said to be a *wavelet* if the system of functions $\{\psi_{j,k} : j,k \in \mathbb{Z}\}$ forms an orthonormal basis for $L^2(\mathbb{R})$, where

$$\psi_{j,k}(x) = 2^{j/2}\psi(2^j x - k), \quad j,k \in \mathbb{Z}.$$

A function $\psi \in L^2(\mathbb{R})$ is a wavelet of $L^2(\mathbb{R})$ if and only if it satisfies the following four conditions:

$$\sum_{j\in\mathbb{Z}} |\hat{\psi}(2^j\xi)|^2 = 1 \quad \text{for a.e. } \xi \in \mathbb{R}. \tag{1}$$

$$\sum_{j\geq 0} \hat{\psi}(2^j\xi)\overline{\hat{\psi}(2^j(\xi+2m\pi))} = 0 \quad \text{for a.e. } \xi \in \mathbb{R} \text{ and for all } m \in 2\mathbb{Z}+1. \tag{2}$$

$$\sum_{k\in\mathbb{Z}} |\hat{\psi}(\xi+2k\pi)|^2 = 1 \quad \text{for a.e. } \xi \in \mathbb{R}. \tag{3}$$

$$\sum_{k\in\mathbb{Z}} \hat{\psi}(\xi+2k\pi)\overline{\hat{\psi}(2^j(\xi+2k\pi))} = 0 \quad \text{for a.e. } \xi \in \mathbb{R} \text{ and for all } j \geq 1. \tag{4}$$

We use the following definition of the Fourier transform

$$\hat{f}(\xi) = \int_{\mathbb{R}} f(x)e^{-i\xi x}dx, \quad \xi \in \mathbb{R}.$$

The fact that equations (1)–(4) characterize all wavelets of $L^2(\mathbb{R})$ was observed by many authors. Lemarié and Meyer [7] obtained these equations for compactly supported wavelets. Bonami, Soria and Weiss [1] proved this fact under the assumption that ψ is band-limited, i.e., $\hat{\psi}$ is compactly supported. For a function ψ to be a wavelet, it is necessary that $\|\psi\|_2 = 1$. With this assumption on ψ, the following theorem was proved independently by G. Gripenberg [2] and X. Wang [9] (see also [3]).

Theorem 1. *Let $\psi \in L^2(\mathbb{R})$ with $\|\psi\|_2 = 1$. Then ψ is a wavelet of $L^2(\mathbb{R})$ if and only if ψ satisfies* (1) *and* (2).

For a proof of this theorem we refer to Chapter 7 of [4]. It was observed in [1] that the orthonormality of the system $\{\psi_{j,k} : j,k \in \mathbb{Z}\}$ is characterized by (3) and (4), and completeness by (1) and (2) (see also [4]). A consequence of Theorem 1 is that equations (1) and (2), along with the assumption $\|\psi\|_2 = 1$, imply the equations (3) and (4).

If ψ is a wavelet, then the support of $\hat{\psi}$ must have measure at least 2π. This minimal measure is achieved if and only if $|\hat{\psi}|$ is the characteristic function of some measurable subset K of $\mathbb{R}$. Such a wavelet is called a *minimally supported frequency* (MSF) wavelet and the associated set K is called a *wavelet set*. We refer to [4] for proofs of the above statements.

A function is said to be *band-limited* if its Fourier transform has compact support. The simplest example of a band-limited wavelet is the Shannon wavelet whose Fourier transform is the characteristic function of the set

$S = [-2\pi, -\pi] \cup [\pi, 2\pi]$. Lemarié and Meyer ([6], [8]) constructed a band-limited wavelet belonging to the Schwartz class. This wavelet satisfies $\hat{\psi}(\xi) = e^{i\xi/2} b(\xi)$, where b is an even, non-negative "bell-shaped" function and its support is equal to the set $[-\frac{8}{3}\pi, -\frac{2}{3}\pi] \cup [\frac{2}{3}\pi, \frac{8}{3}\pi]$. In [1] the authors characterized all wavelets ψ such that $|\hat{\psi}|$ is an even continuous function whose support is equal to the set $[-2\pi - \epsilon', -\pi + \epsilon] \cup [\pi - \epsilon, 2\pi + \epsilon']$, where $\epsilon, \epsilon' > 0$ and $\epsilon + \epsilon' \leq \pi$, so that this class includes the Lemarié-Meyer wavelet as a particular case. Later, in [5] (see also Theorem 4.1, Chapter 3 of [4]), the restriction on ψ (i.e., $|\hat{\psi}|$ is even and continuous) was removed and all wavelets with Fourier transform supported in $[-\frac{8}{3}\pi, -\frac{2}{3}\pi] \cup [\frac{2}{3}\pi, \frac{8}{3}\pi]$ were characterized. (In fact, they characterized all wavelets with Fourier transform supported in the set $[-\frac{8}{3}a, 4\pi - \frac{4}{3}a]$, $0 < a \leq \pi$. It was also shown that for such a wavelet ψ, it is necessary that $\hat{\psi} = 0$ a.e. on $[-\frac{2}{3}a, 2\pi - \frac{4}{3}a]$.) In this article, we generalize this characterization: we construct a set S_n for each $n \geq 2$, so that S_2 is the set $[-\frac{8}{3}\pi, -\frac{2}{3}\pi] \cup [\frac{2}{3}\pi, \frac{8}{3}\pi]$, and characterize all wavelets whose Fourier transform is supported in S_n.

The article is organized as follows. In section 2 we construct a set S_n for each $n \geq 2$, which is symmetric with respect to the origin and is a union of four intervals. In section 3 all wavelets ψ with Fourier transform supported in S_n are characterized, and a necessary and sufficient condition is given for $|\hat{\psi}|$ to be an even function. Next, in section 4 we show that none of these wavelets is associated with an MRA. For a wavelet to be associated with an MRA, it is necessary and sufficient that the corresponding dimension function is equal to 1 a.e. In section 5 we compute the dimension functions explicitly and show that each of these functions is even and assumes all integral values from 0 to the maximum value it attains. This gives an alternative proof of the fact that these wavelets are not MRA-wavelets, and in addition, provides an example of a family of dimension functions.

2 Construction of the set S_n

Let $n \geq 2$. Put

$$\begin{aligned} a_n &= \tfrac{2^{n-1}}{2^n-1}\pi, & b_n &= 2a_n = \tfrac{2^n}{2^n-1}\pi, \\ c_n &= \tfrac{2^{n-1}(2^n-2)}{2^n-1}\pi, & d_n &= 2^n a_n = \tfrac{2^{2n-1}}{2^n-1}\pi. \end{aligned}$$

Define $S_n = S_n^+ \cup S_n^-$, where $S_n^+ = [a_n, b_n] \cup [c_n, d_n]$ and $S_n^- = -(S_n^+)$.

Remark 1.

1. If $n = 2$, then $a_2 = \frac{2}{3}\pi$, $b_2 = c_2 = \frac{4}{3}\pi$ and $d_2 = \frac{8}{3}\pi$. So S_2 is the set associated with the Lemarié-Meyer wavelets.
2. As n increases, the sets $[a_n, b_n]$ move closer to 0 whereas $[c_n, d_n]$ move farther away. Also observe that the measure of the set S_n is $2\big(\frac{2^{n-1}}{2^n-1}\pi + \frac{2^n}{2^n-1}\pi\big)$ which approaches 3π as n tends to infinity.

All terms in the equations (1)–(4) that characterize all wavelets of $L^2(\mathbb{R})$ involve dilations by powers of 2 and translations by integral multiples of 2π. In

the following lemma, for various subsets of S_n, we identify those translates and dilates which are possibly in S_n.

Lemma 1. *For $n \geq 3$, let S_n be as above, and let $e_n = \frac{2^n-2}{2^n-1}\pi$.*

(i) If $\xi \in [a_n, e_n]$, then $\xi + 2k\pi \in S_n$ iff $k = 0$, and $2^j\xi \in S_n$ iff $j = 0$.

(ii) If $\xi \in [e_n, b_n]$, then $\xi + 2k\pi \in S_n$ iff $k = 0, -1$, and $2^j\xi \in S_n$ iff $j = 0, n-1$.

(iii) If $\xi \in [c_n, d_n]$, then $\xi + 2k\pi \in S_n$ iff $k = 0, -2^{n-1}$, and $2^j\xi \in S_n$ iff $j = 0, -(n-1)$.

(iv) If $\xi \in [-e_n, -a_n]$, then $\xi + 2k\pi \in S_n$ iff $k = 0$, and $2^j\xi \in S_n$ iff $j = 0$.

(v) If $\xi \in [-b_n, -e_n]$, then $\xi + 2k\pi \in S_n$ iff $k = 0, 1$, and $2^j\xi \in S_n$ iff $j = 0, n-1$.

(vi) If $\xi \in [-d_n, -c_n]$, then $\xi + 2k\pi \in S_n$ iff $k = 0, 2^{n-1}$, and $2^j\xi \in S_n$ iff $j = 0, -(n-1)$.

Observe that

$$[e_n, b_n] - 2\pi = [-b_n, -e_n],\ 2^{n-1}[e_n, b_n] = [c_n, d_n], \tag{5}$$

and

$$2^{n-1}\big([e_n, b_n] - 2\pi\big) = 2^{n-1}[-b_n, -e_n] = [-d_n, -c_n]. \tag{6}$$

Proof. Let $\xi \in [a_n, e_n]$. If $1 \leq k \leq 2^{n-2} - 1$ then $\xi + 2k\pi \in [b_n, c_n]$, and if $k \geq 2^{n-2}$ then $\xi + 2k\pi \geq d_n$. In either case, $\xi + 2k\pi \notin \text{supp}\ \hat{\psi}$. Similarly, if $-2^{n-2} \leq k \leq -1$, then $\xi + 2k\pi \in [-c_n, -b_n]$, and if $k \leq -2^{n-2} - 1$, then $\xi + 2k\pi \leq -d_n$. Thus, we get $\xi + 2k\pi \in S_n$ if and only if $k = 0$.

Now, let $\xi \in [e_n, b_n]$. In this case, if $1 \leq k \leq 2^{n-2} - 1$ then $\xi + 2k\pi \in [b_n, c_n]$, and $\xi + 2k\pi \geq d_n$ if $k \geq 2^{n-2}$. Also, if $-2^{n-2} \leq k \leq -2$, then $\xi + 2k\pi \in [-c_n, -b_n]$, and if $k \leq -2^{n-2} - 1$, then $\xi + 2k\pi \leq -d_n$. This leaves $k = -1$, in which case $\xi + 2k\pi \in [-b_n, -e_n]$ which is a subset of S_n. Thus, $\xi + 2k\pi \in S_n$ if and only if $k = 0, -1$.

Finally, let $\xi \in [c_n, d_n]$. If $k \leq 1$ then $\xi + 2k\pi \geq d_n$. If $-2^{n-2} + 1 \leq k \leq -1$, then $\xi + 2k\pi \in [b_n, c_n]$. Now, $\xi + 2k\pi \in [-a_n, a_n]$ if $k = 2^{n-2}$, and $\xi + 2k\pi \in [-c_n, -b_n]$ if $-2^{n-1} + 1 \leq k \leq -2^{n-2} - 1$. Also, $\xi + 2k\pi \in [-d_n, -c_n]$ if $k = -2^{n-1}$, and if $k \leq -2^{n-1} - 1$ then $\xi + 2k\pi \leq -d_n$. Thus, in this case, $\xi + 2k\pi \in S_n$ if and only if $k = 0, -2^{n-2}$.

We have proved the statements for the translations in (i)–(iii) of the lemma. The proof for the dilations is similar, and (iv)–(vi) follow by the symmetry of the set S_n. □

3 The main result

The following theorem characterizes all wavelets ψ of $L^2(\mathbb{R})$ such that the support of $\hat{\psi}$ is contained in the set S_n. The case $n = 2$ is Theorem 4.1 of Chapter 3 in [4] in which case the wavelets are MRA-wavelets.

Theorem 2. *Let $n \geq 2$, $\psi \in L^2(\mathbb{R})$, supp $\hat{\psi} \subseteq S_n$ and $b(\xi) = |\hat{\psi}(\xi)|$. Then ψ is a wavelet for $L^2(\mathbb{R})$ if and only if*

(i) $b(\xi) = 1$ *for a.e.* $\xi \in [a_n, e_n] \cup [-e_n, -a_n]$,

(ii) $b^2(\xi) + b^2(2^{n-1}\xi) = 1$ *for a.e.* $\xi \in [e_n, b_n]$,

(iii) $b^2(\xi) + b^2(\xi - 2\pi) = 1$ *for a.e.* $\xi \in [e_n, b_n]$,

(iv) $b(\xi) = b\left(2^{n-1}(\xi - 2\pi)\right)$ *for a.e.* $\xi \in [e_n, b_n]$,

(v) $\hat{\psi}(\xi) = e^{i\theta(\xi)} b(\xi)$, *where θ satisfies*

$$\theta(\xi) + \theta\left(2^{n-1}(\xi - 2\pi)\right) - \theta(\xi - 2\pi) - \theta(2^{n-1}\xi) = (2m(\xi) + 1)\,\pi,$$

for some $m(\xi) \in \mathbb{Z}$, for a.e. $\xi \in [e_n, b_n] \cap (\text{supp } b) \cap (\frac{1}{2^{n-1}}\text{supp } b)$.

Moreover, if $n \geq 3$, then none of these wavelets is associated with an MRA.

Remark 2. 1. It follows from Theorem 2 that $|\hat{\psi}|$ is completely determined by its values on $[e_n, b_n]$. On this set, let $b = |\hat{\psi}|$ be an arbitrary measurable function taking values between 0 and 1. Using equations (5) and (6), $|\hat{\psi}|$ can be extented to other sets of S_n with the help of properties (i)–(iv). Properties (ii), (iii) and (iv) can also be written as

$$b^2(\xi) + b^2\Big(\frac{1}{2^{n-1}}\xi\Big) = 1 \quad \text{for a.e. } \xi \in [c_n, d_n], \tag{7}$$

$$b^2(\xi) + b^2(\xi + 2\pi) = 1 \quad \text{for a.e. } \xi \in [-b_n, -e_n], \tag{8}$$

$$b\left(\frac{1}{2^{n-1}}\xi + 2\pi\right) = b(\xi) \quad \text{for a.e. } [-d_n, -c_n]. \tag{9}$$

2. If $(\text{supp } b) \cap (\frac{1}{2^{n-1}}\text{supp } b)$ has an empty interior in $[e_n, b_n]$, then θ can be chosen to be any measurable function. In particular, we can take $\theta(\xi) = 0$.

Proof of Theorem 2. First let us assume that ψ is a wavelet for $L^2(\mathbb{R})$, supp $\hat{\psi} \subseteq S_n$ and $b = |\hat{\psi}|$. So ψ satisfies the equations (1)–(4). Consider equation (3):

$$\sum_{k\in\mathbb{Z}} b^2(\xi + 2k\pi) = 1 \quad \text{for a.e. } \xi \in \mathbb{R}.$$

We use Lemma 1 to pick out the non-zero terms in this sum. If $|\xi| \in [a_n, e_n]$, then only $k = 0$ will contribute to the sum. So we get $b(\xi) = 1$ a.e., which is (i) of the theorem. For $\xi \in [e_n, b_n]$, we get non-zero contributions from $k = 0$ and $k = -1$. That is, $b^2(\xi) + b^2(\xi - 2\pi) = 1$ for a.e. ξ, which proves (iii). Also if $\xi \in [e_n, b_n]$, then $2^{n-1}\xi \in [c_n, d_n]$, and we get (non-zero terms now correspond to $k = 0$ and $k = -2^{n-1}$)

$$b^2(2^{n-1}\xi) + b^2(2^{n-1}\xi - 2^n\pi) = 1 \quad \text{for a.e. } \xi \in [e_n, b_n]. \tag{10}$$

This can also be written as

$$b^2(\xi) + b^2(\xi - 2^n\pi) = 1 \quad \text{for a.e. } \xi \in [c_n, d_n]. \tag{11}$$

Now consider equation (1):

$$\sum_{j\in\mathbb{Z}} b^2(2^j\xi) = 1 \quad \text{for a.e. } \xi \in \mathbb{R}.$$

For $\xi \in [e_n, b_n]$, the only possible non-zero terms correspond to $j = 0$ and $j = n-1$. So we conclude $b^2(\xi) + b^2(2^{n-1}\xi) = 1$. This proves (ii). Combining (ii) and (10) we get condition (iv).

It remains to prove condition (v). For this purpose we consider equation (4) with $j = n-1$:

$$\sum_{k\in\mathbb{Z}} \hat{\psi}(\xi + 2k\pi)\overline{\hat{\psi}\left(2^{n-1}(\xi + 2k\pi)\right)} = 0 \quad \text{for a.e. } \xi \in \mathbb{R}.$$

If $\xi \in [e_n, b_n]$, the non-zero terms in the sum come from $k = 0$ and -1. Hence,

$$\hat{\psi}(\xi)\overline{\hat{\psi}(2^{n-1}\xi)} + \hat{\psi}(\xi - 2\pi)\overline{\hat{\psi}\left(2^{n-1}(\xi - 2\pi)\right)} = 0 \quad \text{for a.e. } \xi \in [e_n, b_n]. \tag{12}$$

Using (iii), (10) and (12), we see that for almost every $\xi \in [e_n, b_n]$, the vectors

$$\left(\hat{\psi}(\xi), \hat{\psi}(\xi - 2\pi)\right) \quad \text{and} \quad \left(\hat{\psi}(2^{n-1}\xi), \hat{\psi}(2^{n-1}(\xi - 2\pi))\right)$$

are orthonormal in $\mathbb{C}^2$. If we let $\hat{\psi}(\xi) = e^{i\theta(\xi)}b(\xi)$, then it follows that, for some real-valued measurable function α,

$$\begin{aligned} &e^{i\alpha(\xi)}\left(e^{i\theta(\xi)}b(\xi), e^{i\theta(\xi-2\pi)}b(\xi - 2\pi)\right) \\ &\quad = \left(-e^{-i\theta(2^{n-1}(\xi-2\pi))}b(2^{n-1}(\xi - 2\pi)), e^{-i\theta(2^{n-1}\xi)}b(2^{n-1}\xi)\right) \end{aligned} \tag{13}$$

for a.e. $\xi \in [e_n, b_n]$. But from (ii), (iii) and (iv), we know that

$$\left.\begin{array}{rcl} b(2^{n-1}\xi) &=& b(\xi - 2\pi) \\ b(\xi) &=& b\left(2^{n-1}(\xi - 2\pi)\right) \end{array}\right\} \quad \text{for a.e } \xi \in [e_n, b_n]. \tag{14}$$

So (13) can be written as

$$\left.\begin{array}{rcl} \left[e^{i\alpha(\xi)}e^{i\theta(\xi)} + e^{-i\theta\left(2^{n-1}(\xi-2\pi)\right)}\right] b(\xi) &=& 0 \\ \left[e^{i\alpha(\xi)}e^{i\theta(\xi-2\pi)} - e^{-i\theta(2^{n-1}\xi)}\right] b(2^{n-1}\xi) &=& 0 \end{array}\right\} \quad \text{for a.e. } \xi \in [e_n, b_n].$$

This shows that

$$\begin{array}{rcll} e^{-i\alpha(\xi)} &=& e^{i[\theta(\xi)+\theta(2^{n-1}(\xi-2\pi))+\pi]} & \text{for a.e. } \xi \in [e_n, b_n] \cap \text{supp } b. \\ e^{-i\alpha(\xi)} &=& e^{i[\theta(\xi-2\pi)+\theta(2^{n-1}\xi)]} & \text{for a.e. } \xi \in [e_n, b_n] \cap (\frac{1}{2^{n-1}}\text{supp } b). \end{array}$$

Hence, for almost every $\xi \in [e_n, b_n] \cap \text{supp } b \cap (\frac{1}{2^{n-1}}\text{supp } b)$, we have

$$\theta(\xi) + \theta\left(2^{n-1}(\xi - 2\pi)\right) - \theta(\xi - 2\pi) - \theta(2^{n-1}\xi) = (2m(\xi) + 1)\,\pi,$$

for some integer-valued measurable function m. This proves (v).

We now prove the converse. Suppose $\psi \in L^2(\mathbb{R})$, supp $\hat{\psi} \subseteq S_n$ and the function $b(\xi) = |\hat{\psi}(\xi)|$ satisfies conditions (i)–(v) of the theorem. By Theorem 1, to show that ψ is a wavelet, it is sufficient to show that $\|\psi\|_2 = 1$ and ψ satisfies (1) and (2). We have,

$$\begin{aligned} 2\pi\|\psi\|_2^2 &= \|\hat{\psi}\|_2^2 = \int_{S_n} |\hat{\psi}(\xi)|^2\, d\xi \\ &= \left(\int_{-d_n}^{-c_n} + \int_{-b_n}^{-e_n} + \int_{-e_n}^{-a_n} + \int_{a_n}^{e_n} + \int_{e_n}^{b_n} + \int_{c_n}^{d_n}\right) b^2(\xi)\, d\xi \\ &= \int_{-d_n}^{-c_n} + \int_{c_n}^{d_n} + \int_{-b_n}^{-e_n} + \int_{e_n}^{b_n} + 2(e_n - a_n). \end{aligned}$$

By changing variables $\xi \to 2^{n-1}(\xi - 2\pi)$, $\xi \to 2^{n-1}\xi$ and $\xi \to \xi - 2\pi$ in the first, second and third integrals respectively, we get

$$\begin{aligned} 2\pi\|\psi\|_2^2 &= 2^{n-1}\int_{e_n}^{b_n} \left[b^2\left(2^{n-1}(\xi - 2\pi)\right) + b^2(2^{n-1}\xi)\right] d\xi \\ &\quad + \int_{e_n}^{b_n} \left[b^2(\xi - 2\pi) + b^2(\xi)\right] d\xi + 2(e_n - a_n) \\ &= 2^{n-1}(b_n - e_n) + (b_n - e_n) + 2(e_n - a_n) = 2\pi, \end{aligned}$$

where we have used (10) and property (iii) of the theorem. Hence, $\|\psi\|_2 = 1$.

We will now show that ψ satisfies (1). Let

$$\rho(\xi) = \sum_{j \in \mathbb{Z}} |\, \hat{\psi}(2^j \xi)\, |^2\, .$$

Suppose $\xi > 0$. Observe that

$$\begin{aligned} \mathbb{R}^+ &= \bigcup_{l \in \mathbb{Z}} 2^l [a_n, b_n] \\ &= \bigcup_{l \in \mathbb{Z}} 2^l \left([a_n, e_n] \cup [e_n, b_n]\right). \end{aligned}$$

So there is an $l \in \mathbb{Z}$ such that $\xi \in 2^l[a_n, e_n] \cup 2^l[e_n, b_n]$. If $\xi \in 2^l[a_n, e_n]$ then since $2^{-l}\xi \in [a_n, e_n]$, by Lemma 1 we have, $2^j(2^{-l}\xi) \notin S_n$ if $j \neq 0$. That is, $2^j\xi \notin S_n$ if $j \neq -l$. Hence, $\rho(\xi) = 1$, by (i). Similarly by using (ii), we can prove that $\rho(\xi) = 1$ if $\xi \in 2^l[e_n, b_n]$. A similar decomposition for $\xi < 0$ proves that $\rho(\xi) = 1$ for a.e. $\xi \in \mathbb{R}$.

Finally, we have to show that ψ satisfies (2). For $m \in 2\mathbb{Z}+1$, let us denote the function on the left hand side of (2) by $t_m(\xi)$. Then

$$\begin{aligned} t_m(\xi) &= \sum_{j\geq 0} \hat{\psi}(2^j\xi)\overline{\hat{\psi}\left(2^j(\xi+2m\pi)\right)} \\ &= \sum_{j\geq 0} \hat{\psi}\left(2^j(\xi+2m\pi-2m\pi)\right)\overline{\hat{\psi}\left(2^j(\xi+2m\pi)\right)} \\ &= \overline{t_{-m}(\xi+2m\pi)}. \end{aligned} \tag{15}$$

Therefore, we have only to show that $t_m(\xi)=0$ for a.e. ξ, if $m \in 2\mathbb{Z}^- + 1$ (negative odd integers). Suppose $m \in 2\mathbb{Z}^- + 1$ and $m \neq -1$. Let $\xi \in \mathbb{R}$ and suppose that $2^j\xi \in S_n$. Since $j \geq 0$, $n \geq 2$ and $m \neq -1$ and is odd, therefore $2^j m \neq 0, \pm 1, \pm 2^{n-1}$. So by Lemma 1, we have $2^j\xi + 2\cdot 2^j m\pi \notin S_n$. This implies that each term of $t_m(\xi)$ is zero, which proves that t_m is zero. It now remains to show that $t_{-1}(\xi)=0$ for a.e ξ. We have

$$t_{-1}(\xi) = \sum_{j\geq 0} \hat{\psi}(2^j\xi)\overline{\hat{\psi}(2^j\xi - 2\cdot 2^j\pi)}.$$

Let $\xi \in \mathbb{R}$ and suppose that $2^j\xi \in S_n$. Then again by Lemma 1, $2^j\xi - 2\cdot 2^j\pi \notin S_n$ if $-2^j \neq 0, \pm 1, \pm 2^{n-1}$. So the only possible j's to contribute a non-zero term are $j=0$ and $j=n-1$. Thus,

$$t_{-1}(\xi) = \hat{\psi}(\xi)\overline{\hat{\psi}(\xi-2\pi)} + \hat{\psi}(2^{n-1}\xi)\overline{\hat{\psi}(2^{n-1}\xi - 2\cdot 2^{n-1}\pi)}. \tag{16}$$

Now, both ξ and $\xi - 2\pi$ belong to S_n only if $\xi \in [e_n, b_n]$. Hence, the first term of (16) is zero unless $\xi \in [e_n, b_n]$. Similarly, both $2^{n-1}\xi$ and $2^{n-1}\xi - 2\cdot 2^{n-1}\pi$ belong to S_n only when $2^{n-1}\xi \in [c_n, d_n]$, which is equivalent to saying that $\xi \in [e_n, b_n]$. That is, the second term of (16) is also zero unless $\xi \in [e_n, b_n]$. Thus we get $t_{-1}(\xi)=0$ if $\xi \notin [e_n, b_n]$. Now on $[e_n, b_n]$, we have, by (14)

$$t_{-1}(\xi) = b(\xi)b(2^{n-1}\xi)e^{i[\theta(\xi)-\theta(\xi-2\pi)]} + b(2^{n-1}\xi)b(\xi)e^{i[\theta(2^{n-1}\xi)+\theta\left(2^{n-1}(\xi-2\pi)\right)]}.$$

If $\xi \notin [e_n, b_n] \cap \operatorname{supp} b \cap (\frac{1}{2^{n-1}}\operatorname{supp} b)$, then either $b(\xi)=0$ or $b(2^{n-1}\xi)=0$. So $t_{-1}(\xi)=0$. And if $\xi \in [e_n, b_n] \cap \operatorname{supp} b \cap (\frac{1}{2^{n-1}}\operatorname{supp} b)$, then by (v), $t_{-1}(\xi)=0$. This completes the characterization. □

There is another equation which characterizes all wavelets ψ such that $\hat{\psi}$ is supported in the set S_n and the function $b = |\hat{\psi}|$ is even. For $n=2$, the following proposition is proved in [4] (Proposition 4.7, Chapter 3). We observe that the result can be extended to the general case.

Proposition 1. *Suppose that ψ is a wavelet of $L^2(\mathbb{R})$, $b = |\hat{\psi}|$, and* supp $b \subseteq S_n$. *Then b is almost everywhere even if and only if*

$$b^2(\xi) + b^2(2\pi - \xi) = 1 \quad \text{for a.e. } \xi \in [e_n, b_n]. \tag{17}$$

Proof. Let ψ be a wavelet of $L^2(\mathbb{R})$ such that $b = |\hat{\psi}|$ is supported in S_n. Suppose that b is an even function and $\xi \in [e_n, b_n]$. Since $-\xi \in [-b_n, -e_n]$, we have

$$\begin{aligned} 1 &= b^2(-\xi) + b^2(-\xi + 2\pi), \quad \text{by (8)} \\ &= b^2(\xi) + b^2(2\pi - \xi), \quad \text{since } b \text{ is even,} \end{aligned}$$

which is (17).

Conversely, suppose that (17) holds. Since by (i) of Theorem 2, b is even on $[a_n, e_n]$, it is enough to show that b is even on the sets $[e_n, b_n]$ and $[c_n, d_n]$.

(i) Let $\xi \in [e_n, b_n]$. Therefore, $-\xi \in [-b_n, -e_n]$. Then

$$b^2(-\xi) + b^2(-\xi + 2\pi) = 1, \quad \text{by (8).}$$

This fact, together with (17), gives us

$$b(\xi) = b(-\xi). \tag{18}$$

(ii) Now let $\xi \in [c_n, d_n]$. Therefore, $\frac{1}{2^{n-1}}\xi \in [e_n, b_n]$. From (18) we get

$$b\Big(\frac{1}{2^{n-1}}\xi\Big) = b\Big(-\frac{1}{2^{n-1}}\xi\Big). \tag{19}$$

Since $-\frac{1}{2^{n-1}}\xi \in [-b_n, -e_n]$, using (8) we obtain

$$b^2\Big(-\frac{1}{2^{n-1}}\xi\Big) + b^2\Big(-\frac{1}{2^{n-1}}\xi + 2\pi\Big) = 1. \tag{20}$$

Now, since $-\xi \in [-d_n, -c_n]$, we get (using (9))

$$b\Big(-\frac{1}{2^{n-1}}\xi + 2\pi\Big) = b(-\xi). \tag{21}$$

Substituting (19) and (21) in (20), we get

$$b^2\Big(\frac{1}{2^{n-1}}\xi\Big) + b^2(-\xi) = 1. \tag{22}$$

Comparing (22) and (7) we get $b(\xi) = b(-\xi)$, which proves the proposition. □

4 The wavelets associated with S_n, $n \geq 3$ are non-MRA

A multiresolution analysis (MRA) is a sequence of closed subspaces $\{V_j : j \in \mathbb{Z}\}$ of $L^2(\mathbb{R})$ satisfying the following properties:

(i) $V_j \subset V_{j+1}$ for all $j \in \mathbb{Z}$

(ii) $\cup_{j\in\mathbb{Z}} V_j$ is dense in $L^2(\mathbb{R})$ and $\cap_{j\in\mathbb{Z}} V_j = \{0\}$

(iii) $f \in V_j$ if and only if $f(2\cdot) \in V_{j+1}$ for all $j \in \mathbb{Z}$

(iv) there exists a function $\varphi \in L^2(\mathbb{R})$ such that $\{\varphi(\cdot - k) : k \in \mathbb{Z}\}$ forms an orthonormal basis for V_0.

A function φ that satisfies property (iv) is called a *scaling function* for the MRA. If φ is a scaling function for a given MRA, then it is easy to see that (see [4]) there exists a 2π-periodic function m_0 in $L^2(\mathbb{T})$, called the *low-pass filter*, such that

$$\hat{\varphi}(2\xi) = m_0(\xi)\hat{\varphi}(\xi). \tag{23}$$

The function m_0 satisfies the equation

$$|m_0(\xi)|^2 + |m_0(\xi + \pi)|^2 = 1 \quad \text{for a.e } \xi \in \mathbb{T}. \tag{24}$$

Using (23) and (24), one can then construct a wavelet ψ associated with the MRA. For any such wavelet the following equation holds (see [4] for details):

$$|\hat{\varphi}(\xi)|^2 = \sum_{j\geq 1} |\hat{\psi}(2^j\xi)|^2. \tag{25}$$

We will now prove the last statement in Theorem 2.

Proposition 2. *If $n \geq 3$, then the wavelets characterized in Theorem 2 are not associated with any MRA.*

Proof. Let ψ be any wavelet such that $\hat{\psi}$ is supported in S_n and let it be associated with an MRA. Let φ be the corresponding scaling function and m_0 be the low-pass filter associated with φ. Using (25) we can easily find that

$$|\hat{\varphi}(\xi)| = \begin{cases} 1 & \text{if } |\xi| \leq a_n \\ b(2^{n-l-1}\xi) & \text{if } |\xi| \in 2^l[e_n, b_n],\ 0 \leq l \leq n-2 \\ 0 & \text{otherwise.} \end{cases} \tag{26}$$

Hence,

$$|\hat{\varphi}(2\xi)| = \begin{cases} 1 & \text{if } |\xi| \leq \frac{a_n}{2} \\ b(2^{n-l}\xi) & \text{if } |\xi| \in 2^{l-1}[e_n, b_n],\ 0 \leq l \leq n-2 \\ 0 & \text{otherwise.} \end{cases} \tag{27}$$

Case 1. $b(\xi) \not\equiv 1$ on $[e_n, b_n]$, i.e., $b(\xi) \not\equiv 0$ on $[c_n, d_n]$.

If $|\xi| \leq a_n$, then from (23) and (26), we have $|\hat{\varphi}(2\xi)| = |m_0(\xi)| \cdot |\hat{\varphi}(\xi)| = |m_0(\xi)|$. Therefore, by (27), we obtain $|m_0(\xi)| = 1$ if $|\xi| \leq \frac{a_n}{2}$. Since m_0 is 2π-periodic, we have

$$|m_0(\xi)| = 1 \quad \text{on } \left[-\tfrac{a_n}{2}, \tfrac{a_n}{2}\right] + 2 \cdot 2^{n-3}\pi.$$

Note that $[-\frac{a_n}{2}, \frac{a_n}{2}] + 2 \cdot 2^{n-3}\pi = [\frac{c_n}{2}, \frac{d_n}{2}] = 2^{n-2}[e_n, b_n]$. Now, from (26), on $2^{n-2}[e_n, b_n]$ we have $|\hat{\varphi}(\xi)| = b(2\xi)$. So $|\hat{\varphi}(2\xi)| = |m_0(\xi)| \cdot |\hat{\varphi}(\xi)| = b(2\xi)$. But

by (27), $|\hat{\varphi}(2\xi)| = 0$ on $2^{n-2}[e_n, b_n]$. Therefore, $b(2\xi) = 0$ on $2^{n-2}[e_n, b_n]$. That is, $b(\xi) = 0$ on $2^{n-1}[e_n, b_n] = [c_n, d_n]$, which is a contradiction.

Case 2. $b(\xi) \equiv 1$ on $[e_n, b_n]$.

As above, we have

$$|\hat{\varphi}(\xi)| = \begin{cases} 1 & \text{if } |\xi| \le a_n \\ & \text{or if } \xi \in 2^l[-b_n, -e_n],\ 0 \le l \le n-2 \\ 0 & \text{otherwise.} \end{cases} \tag{28}$$

Therefore,

$$|\hat{\varphi}(2\xi)| = \begin{cases} 1 & \text{if } |\xi| \le \frac{a_n}{2} \\ & \text{or if } \xi \in 2^{l-1}[-b_n, -e_n],\ 0 \le l \le n-2 \\ 0 & \text{otherwise.} \end{cases} \tag{29}$$

Now $|\hat{\varphi}(2\xi)| = |m_0(\xi)| \cdot |\hat{\varphi}(\xi)| = |m_0(\xi)|$ if $|\xi| \le a_n$, by (28). Therefore, by (29), $|m_0(\xi)| = 1$ if $|\xi| \le \frac{a_n}{2}$. Using the 2π-periodicity of m_0, we get

$$|m_0(\xi)| = 1 \quad \text{on } \left[-\tfrac{a_n}{2}, \tfrac{a_n}{2}\right] - 2 \cdot 2^{n-3}\pi.$$

Note that $[-\frac{a_n}{2}, \frac{a_n}{2}] - 2 \cdot 2^{n-3}\pi = [-\frac{d_n}{2}, -\frac{c_n}{2}] = 2^{n-2}[-b_n, -e_n]$. Hence, on the interval $2^{n-2}[-b_n, -e_n]$, we have $|\hat{\varphi}(2\xi)| = |m_0(\xi)| \cdot |\hat{\varphi}(\xi)| = |m_0(\xi)| = 1$. But by (29), $|\hat{\varphi}(2\xi)| = 0$ on $2^{n-2}[-b_n, -e_n]$, which is again a contradiction.

Therefore, for $n \ge 3$, the wavelets characterized in Theorem 2 are non-MRA wavelets. □

5 The dimension functions of wavelets associated with S_n

Let ψ be any wavelet of $L^2(\mathbb{R})$. Define V_j to be the closure of the span of $\{\psi_{l,k} : l < j,\ k \in \mathbb{Z}\}$. If $\{V_j : j \in \mathbb{Z}\}$ forms an MRA of $L^2(\mathbb{R})$, then we say that ψ is associated with an MRA, or ψ is an MRA-wavelet.

Not every wavelet of $L^2(\mathbb{R})$ is associated with an MRA. In the last section we proved that the wavelets associated with the set S_n are non-MRA wavelets. The first example of a non-MRA wavelet, which is an MSF wavelet, was given by J.L. Journé. The associated wavelet set is $J = \left[-\frac{32}{7}\pi, -4\pi\right] \cup \left[-\pi, -\frac{4}{7}\pi\right] \cup \left[\frac{4}{7}\pi, \pi\right] \cup \left[4\pi, \frac{32}{7}\pi\right]$. Another interesting non-MRA wavelet is the Lemarié wavelet which is also an MSF wavelet with the corresponding wavelet set $L = \left[-\frac{8}{7}\pi, -\frac{4}{7}\pi\right] \cup \left[\frac{4}{7}\pi, \frac{6}{7}\pi\right] \cup \left[\frac{24}{7}\pi, \frac{32}{7}\pi\right]$. These two wavelets belong to the class of wavelets characterized in section 3 (see Remark 3).

Given a wavelet ψ of $L^2(\mathbb{R})$, there is an associated function D_ψ, called the *dimension function*, defined by

$$D_\psi(\xi) = \sum_{j \ge 1} \sum_{k \in \mathbb{Z}} |\hat{\psi}\left(2^j(\xi + 2k\pi)\right)|^2.$$

It was proved independently by Gripenberg and Wang that a wavelet ψ is an MRA-wavelet if and only if $D_\psi = 1$ a.e. For a proof of this fact see [4].

In this section we compute the dimension functions for the wavelets characterized in Theorem 2. In addition to providing a family of dimension functions, this will give an alternative proof of the fact that these wavelets are not associated with any MRA. These dimension functions are symmetric with respect to the origin and attain all integral values starting from 0 to the maximum value. Also note that for a fixed n, all wavelets whose Fourier transform is supported in S_n have the same dimension function D_n.

Theorem 3. *Fix $n \geq 3$. Let ψ be a wavelet such that* supp $\hat{\psi} \subseteq S_n$ *and let D_n be the dimension function associated with ψ. Then*

$$D_n(\xi) = \begin{cases} n-1 & \text{a.e. if } |\xi| \in \left[0, \frac{2}{2^n-1}\pi\right] \\ l-1 & \text{a.e. if } |\xi| \in \left[\frac{2^{n-l}}{2^n-1}\pi, \frac{2^{n-l+1}}{2^n-1}\pi\right] \quad (2 \leq l \leq n-1) \\ 0 & \text{a.e. if } |\xi| \in \left[\frac{2^{n-1}}{2^n-1}\pi, \frac{2^n-2}{2^n-1}\pi\right] = [a_n, e_n] \\ 1 & \text{a.e. if } |\xi| \in \left[\frac{2^n-2}{2^n-1}\pi, \pi\right] = [e_n, \pi]. \end{cases}$$

Proof. For $l \in \mathbb{Z}$, we define

$$p_l = \frac{2^{n+l-1}}{2^n-1}\pi = 2^l a_n \quad \text{and} \quad q_l = \frac{2^{n+l-1}-2^l}{2^n-1}\pi = 2^{l-1}e_n.$$

Note that $p_0 = a_n$, $p_1 = 2p_0 = b_n$, $p_n = d_n$, $q_1 = e_n$ and $q_n = c_n$. Also observe that $p_l < q_{l+1} < p_{l+1}$ for all $l \geq 0$. As D_n is 2π-periodic, it is enough to compute its values for $\xi \in [-\pi, \pi]$. Since $[-\pi, \pi] \subset [-p_1, p_1]$, we will compute D_n for the interval $[-p_1, p_1]$. Note that

$$(0, p_1] = \bigcup_{l \geq 0} 2^{-l}[p_0, p_1] = \bigcup_{l \geq 0} [p_{-l}, p_{-l+1}].$$

Case 1. $\xi \in [\frac{2^{n-1}}{2^n-1}\pi, \frac{2^n}{2^n-1}\pi] = [p_0, p_1]$.

An elementary calculation shows that if $k \neq 0, -1$, then $2^j(\xi + 2k\pi)$ lies outside the support of $\hat{\psi}$ for all $j \geq 1$. Therefore, we have only to consider $k = 0, -1$. We consider the intervals $[p_0, q_1]$ and $[q_1, p_1]$ separately.

(a) If $\xi \in [p_0, q_1]$, then $2^j\xi \in [b_n, c_n]$ for $1 \leq j \leq n-1$ and $2^j\xi \geq d_n$ for $j \geq n$. By a similar argument $2^j(\xi - 2\pi) \notin$ supp $\hat{\psi}$ for all $j \geq 1$. So $D_n(\xi) = 0$ for a.e. $\xi \in [p_0, q_1]$.

(b) If $\xi \in [q_1, p_1]$, then it can be shown that $2^j(\xi - 2\pi) \notin$ supp $\hat{\psi}$ if $j \neq n-1$. Therefore, $D_n(\xi) = |\hat{\psi}(2^{n-1}\xi)|^2 + |\hat{\psi}\left(2^{n-1}(\xi - 2\pi)\right)|^2 = 1$, by (11), as $2^{n-1}\xi \in [q_n, p_n] = [c_n, d_n]$.

We now proceed to compute D_n on other sets. Observe that

$$D_n(\xi) \quad = \quad \sum_{k=-\infty}^{-1} \sum_{j \geq 1} \left|\hat{\psi}\left(2^j(\xi + 2k\pi)\right)\right|^2 + \sum_{j \geq 1} \left|\hat{\psi}(2^j\xi)\right|^2 +$$

$$
\begin{aligned}
&+\sum_{k=1}^{\infty}\sum_{j\geq 1}\left|\hat{\psi}\left(2^j(\xi+2k\pi)\right)\right|^2\\
&= \quad D_n^-(\xi)+D_n^0(\xi)+D_n^+(\xi), \text{ say.}
\end{aligned}
$$

Let $\xi \in \left[\frac{2^{n-l}}{2^n-1}\pi, \frac{2^{n-l+1}}{2^n-1}\pi\right] = [p_{-l+1}, p_{-l+2}]$, $l \geq 2$.

In the sum $D_n^-(\xi)$, the only pairs of (j,k) giving a possible non-zero term correspond to $j = n-m-2$ and $k = -2^m$, $0 \leq m \leq n-3$. Thus,

$$D_n^-(\xi) = \sum_{m=0}^{n-3}\left|\hat{\psi}\left(2^{n-m-2}(\xi - 2\cdot 2^m\pi)\right)\right|^2.$$

Similarly,

$$D_n^+(\xi) = \sum_{m=0}^{n-3}\left|\hat{\psi}\left(2^{n-m-2}(\xi + 2\cdot 2^m\pi)\right)\right|^2.$$

Case 2. $\xi \in \left[\frac{2^{n-l}}{2^n-1}\pi, \frac{2^{n-l+1}}{2^n-1}\pi\right] = [p_{-l+1}, p_{-l+2}]$, $l \geq n$.

Observe that in this case

$$2^{n-m-2}(\xi + 2\cdot 2^m\pi) \in [c_n, d_n], \quad \text{for } \ 0 \leq m \leq n-3.$$

Therefore,

$$
\begin{aligned}
&|\hat{\psi}\big(2^{n-m-2}(\xi+2\cdot 2^m\pi)\big)|^2 + |\hat{\psi}\big(2^{n-m-2}(\xi-2\cdot 2^m\pi)\big)|^2\\
= \quad &|\hat{\psi}\big(2^{n-m-2}(\xi+2\cdot 2^m\pi)\big)|^2 + |\hat{\psi}\big(2^{n-m-2}(\xi+2\cdot 2^m\pi) - 2^{n-1}\pi\big)|^2\\
= \quad &1, \quad \text{by (11).}
\end{aligned}
$$

Hence, we get $D_n^+(\xi) + D_n^-(\xi) = n-2$.

Now, $[p_{-l+1}, p_{-l+2}] = [p_{-l+1}, q_{-l+2}] \cup [q_{-l+2}, p_{-l+2}]$.

(a) If $\xi \in [p_{-l+1}, q_{-l+2}]$, then $2^j\xi \notin \text{supp } \hat{\psi}$, if $j \neq l-1$. But $2^{l-1}\xi \in [p_0, q_1] = [a_1, e_n]$. Therefore, $D_n^0(\xi) = |\hat{\psi}(2^{l-1}\xi)|^2 = 1$.

(b) If $\xi \in [q_{-l+2}, p_{-l+2}]$, then $2^j\xi \notin \text{supp } \hat{\psi}$ if $j \neq l-1, n+l-2$. Now $2^{n+l-2}\xi \in [q_n, p_n] = [c_n, d_n]$. By (7) we get $D_n^0(\xi) = |\hat{\psi}(2^{n+l-2}\xi)|^2 + |\hat{\psi}(2^{l-1}\xi)|^2 = 1$.

Thus, in either case $D_n(\xi) = n-2+1 = n-1$. We have proved that $D_n(\xi) = n-1$, if $\xi \in \left[\frac{2^{n-l}}{2^n-1}\pi, \frac{2^{n-l+1}}{2^n-1}\pi\right]$, $l \geq n$. That is,

$$D_n(\xi) = n-1 \text{ for a.e. } \xi \in [0, \tfrac{2}{2^n-1}\pi].$$

Case 3. $\xi \in \left[\frac{2^{n-l}}{2^n-1}\pi, \frac{2^{n-l+1}}{2^n-1}\pi\right]$, $l = 2$.

As in Case 2, it can be shown that

$$2^{n-m-2}(\xi - 2\cdot 2^m\pi) \quad \text{and} \quad 2^{n-m-2}(\xi + 2\cdot 2^m\pi)$$

are not in supp $\hat{\psi}$ for all m such that $0 \le m \le n-3$. Therefore, $D_n^-(\xi) = D_n^+(\xi) = 0$.

Now if $\xi \in \left[\frac{2^{n-2}}{2^n-1}\pi, \frac{2^{n-1}-1}{2^n-1}\pi\right]$, then $2^j\xi \notin$ supp $\hat{\psi}$ if $j \neq 1$; and if ξ belongs to the set $\left[\frac{2^{n-1}-1}{2^n-1}\pi, \frac{2^{n-2}}{2^n-1}\pi\right]$, then $2^j\xi \notin$ supp $\hat{\psi}$ if $j \neq 1, n$. In either case,

$$D_n(\xi) = D_n^0(\xi) = 1.$$

Case 4. $\xi \in \left[\frac{2^{n-l}}{2^n-1}\pi, \frac{2^{n-l+1}}{2^n-1}\pi\right]$, $3 \le l \le n-1$. (This case is required if $n \ge 4$.)

We can write $2^{n-m-2}\xi \in \left[\frac{2^{n+p-l}}{2^n-1}\pi, \frac{2^{n+p}}{2^n-1}\pi\right]$, where $p = n-m-l-1$. Using the condition on support of $\hat{\psi}$, we can show that if $\xi \in \left[\frac{2^{n+p-l}}{2^n-1}\pi, \frac{2^{n+p}}{2^n-1}\pi\right]$, then $\xi \pm 2\cdot 2^{n-2}\pi \in$ supp $\hat{\psi}$ only when $p \le -1$. So in order that $\hat{\psi}(2^{n-m-2}\xi \pm 2\cdot 2^{n-2}\pi) \neq 0$, we must have $p = n-m-l-1 \le -1$, i.e., $m \ge n-l$. We also have, $0 \le m \le n-3$. Therefore, $1 \le n-m-2 \le l-2$. Thus

$$\begin{aligned} D_n(\xi) &= \sum_{j\ge 1} |\hat{\psi}(2^j\xi)|^2 \\ &\quad + \sum_{m=0}^{n-3} \left\{ |\hat{\psi}(2^{n-m-2}(\xi + 2\cdot 2^m\pi))|^2 + \left|\hat{\psi}(2^{n-m-2}(\xi - 2\cdot 2^m\pi))\right|^2 \right\} \\ &= \sum_{j\ge 1} \left|\hat{\psi}(2^j\xi)\right|^2 + \sum_{s=1}^{l-2} \left\{ \left|\hat{\psi}(2^s\xi + 2\cdot 2^{n-2}\pi)\right|^2 + \left|\hat{\psi}(2^s\xi - 2\cdot 2^{n-2}\pi)\right|^2 \right\}. \end{aligned}$$

Now, $2^s\xi + 2\cdot 2^{n-2}\pi \in [c_n, d_n]$, $1 \le s \le l-2$. By (11), the second sum on the right hand side of the last equality above is equal to $l-2$. Also as in the previous cases, the first sum can be shown to be equal to 1. Hence, $D_n(\xi) = l-1$. □

Remark 3. In [3] MSF wavelets were considered where they were called unimodular wavelets. Let ψ be an MSF wavelet and $K =$ supp $\hat{\psi}$.

Define $K^+ = K \cap [0,\infty)$ and $K^- = K \cap (-\infty, 0]$. One of the results proved in [3] is about wavelets ψ such that $K^- = -K^+$ and K^+ consists of two disjoint intervals. They proved the following:

(A) Let $\hat{\psi}(\xi) = \mu(\xi)\chi_K(\xi)$, where $K = K^+ \cup K^-, K^- = -K^+$ and $K^+ = [a_n, \pi] \cup [2^{n-1}\pi, d_n]$ with $|\mu(\xi)| = 1$. Then ψ is a wavelet for $L^2(\mathbb{R})$. Moreover, each MSF wavelet for which $K^- = -K^+$, and K^+ is a union of two disjoint intervals is of this form.

If we define $b = \chi_{[e_n,\pi]}$ on $[e_n, b_n]$, and extend it to S_n by using (i)–(iv) of Theorem 2, then we get all wavelets described in (A). In particular, for $n = 3$, the corresponding wavelet is the Journé wavelet.

The paper [3] also contains the following result.

(B) Let $\hat{\psi}(\xi) = \mu(\xi)\chi_K(\xi)$ where

$$K^- = \left[-2\left(1 - \frac{2p+1}{2^n-1}\right)\pi, -\left(1 - \frac{2p+1}{2^n-1}\right)\pi\right]$$

and

$$K^+ = \left[\frac{2(p+1)}{2^n-1}\pi, \frac{2(2p+1)}{2^n-1}\pi\right] \cup \left[\frac{2^n(2p+1)}{2^n-1}\pi, \frac{2^{n+1}(p+1)}{2^n-1}\pi\right]$$

for $n \geq 3,\ 1 \leq p \leq 2^{n-1}-2$ and $|\mu(\xi)| = 1$. Then ψ is a wavelet for $L^2(\mathbb{R})$. Moreover, each MSF wavelet for which K^- is an interval and K^+ is the union of two disjoint intervals is of this form.

If we define $b(\xi) = 0$ on the interval $[e_n, b_n]$ and extend it to S_n by using (i)–(iv) of Theorem 2, then we get

$$b = \chi_{[-b_n,-a_n]\cup[a_n,e_n]\cup[c_n,d_n]}$$

which corresponds to the case $p = 2^{n-2}-1$ in (B). In particular, for $n = 3$, we get the Lemarié wavelet.

Acknowledgements

The first author was supported by The National Board for Higher Mathematics (NBHM), Govt. of India.

References

[1] A. Bonami, F. Soria, and G. Weiss, Band-limited wavelets, *J. Geom. Anal.* 3, 543-578 (1993).

[2] G. Gripenberg, A necessary and sufficient condition for the existence of a father wavelet, *Studia Math.* 114, 207-226 (1995).

[3] Y. -H. Ha, H. Kang, J. Lee, and J. Seo, Unimodular wavelets for L^2 and the Hardy space H^2, *Michigan Math. J.* 41, 345-361 (1994).

[4] E. Hernández and G. Weiss, *A First Course on Wavelets*, CRC Press, Boca Raton, FL, 1996.

[5] E. Hernández, X. Wang, and G. Weiss, Smoothing minimally supported frequency wavelets: I, *J. Fourier Anal. Appl.* 2, 329-340 (1995).

[6] P. G. Lemarié and Y. Meyer, Ondelettes et bases Hilbertiennes, *Rev. Mat. Iberoamericana.* 2, 1-18 (1986).

[7] P. G. Lemarié and Y. Meyer, Analyse multi-echelles et ondelettes a support compact, in *Les Ondelettes en 1989* (P. G. Lemarié, ed.), Lecture Notes in Math. 1438, 1989, pp. 26-38.

[8] Y. Meyer, Principe d'incertitude, bases hilbertiennes et algèbres d'opérateurs, Séminaire Bourbaki. 662 (1985-86).

[9] X. Wang, The study of wavelets from the properties of their Fourier transforms, Ph.D. thesis, Washington University in St. Louis (1995).

JOURNAL OF CONCRETE AND APPLICABLE MATHEMATICS,VOL.3,NO.1,91-103,2005,COPYRIGHT 2005 EUDOXUS PRESS,LLC

Ostrowski Type Inequalities for Functions Defined on Linear Spaces and Applications for Semi-Inner Products

S.S. Dragomir
School of Computer Science and Mathematics
Victoria University of Technology
PO Box 14428, Melbourne City MC
Victoria 8001, Australia.
Email: sever@csm.vu.edu.au

Abstract

A generalisation of Ostrowski's inequality for functions defined on linear spaces in terms of Gâteaux derivatives is given. Applications for semi-inner products in normed linear spaces are also pointed out.

Keywords: Ostrowski's Inequality, Semi-Inner Products.
AMS Subject Classification Codes: Primary 26D15; Secondary 46B05.

1 Introduction

The following result is known in the literature as Ostrowski's inequality (see e.g. [18]).

Theorem 1 *Let $f : [a,b] \to \mathbb{R}$ be a differentiable mapping on (a,b) with the property that $|f'(t)| \leq M$ for all $t \in (a,b)$. Then*

$$\left| f(x) - \frac{1}{b-a}\int_a^b f(t)\,dt \right| \leq \left[\frac{1}{4} + \frac{\left(x - \frac{a+b}{2}\right)^2}{(b-a)^2} \right] (b-a)\,M, \tag{1}$$

for all $x \in [a,b]$.

The constant $\frac{1}{4}$ is the best possible constant in the sense that it cannot be replaced by a smaller constant.

This result has been generalised for absolutely continuous functions $f : [a,b] \to \mathbb{R}$ (see [19], [20] and [22]) as follows.

Theorem 2 *Let $f:[a,b]\to\mathbb{R}$ be absolutely continuous on $[a,b]$. Then for all $x\in[a,b]$, we have*

$$\left|f(x)-\frac{1}{b-a}\int_a^b f(t)\,dt\right| \tag{2}$$

$$\leq\begin{cases}\left[\frac{1}{4}+\left(\frac{x-\frac{a+b}{2}}{b-a}\right)^2\right](b-a)\|f'\|_\infty, & \text{if } f'\in L_\infty[a,b];\\ \frac{1}{(p+1)^{\frac{1}{p}}}\left[\left(\frac{x-a}{b-a}\right)^{p+1}+\left(\frac{b-x}{b-a}\right)^{p+1}\right]^{\frac{1}{p}}(b-a)^{\frac{1}{p}}\|f'\|_q, & \text{if } f'\in L_q[a,b];\\ \left[\frac{1}{2}+\left|\frac{x-\frac{a+b}{2}}{b-a}\right|\right]\|f'\|_1, & \end{cases}$$

where $\|\cdot\|_r$ $(r\in[1,\infty])$ are the usual Lebesgue norms on $L_r[a,b]$, i.e.,

$$\|g\|_\infty:=ess\sup_{t\in[a,b]}|g(t)|$$

and

$$\|g\|_r:=\left(\int_a^b|g(t)|^r\,dt\right)^{\frac{1}{r}},\quad r\in[1,\infty).$$

The constants $\frac{1}{4}$, $\frac{1}{(p+1)^{\frac{1}{p}}}$ and $\frac{1}{2}$ respectively are sharp in the sense mentioned in Theorem 1.

The above inequalities can also be obtained from Fink's result in [23] on choosing $n=1$ and performing some appropriate computations.

Recently, the Ostrowski type inequalities (2) were extended to functions with values in Banach spaces (see [2]).

Theorem 3 *Let $(X,\|\cdot\|)$ be a Banach space with the Radon-Nikodym property and $f:[a,b]\to X$ an absolutely continuous function on $[a,b]$. Then we have the inequalities*

$$\left\|f(s)-\frac{1}{b-a}(B)\int_a^b f(t)\,dt\right\|$$

$$\leq\begin{cases}\left[\frac{1}{4}+\left(\frac{x-\frac{a+b}{2}}{b-a}\right)^2\right](b-a)\|\|f'\|\|_{[a,b],\infty}, & \text{if } f'\in L_\infty([a,b];X);\\ \frac{1}{(q+1)^{\frac{1}{q}}}\left[\left(\frac{s-a}{b-a}\right)^{q+1}+\left(\frac{b-s}{b-a}\right)^{q+1}\right]^{\frac{1}{q}}(b-a)^{\frac{1}{q}}\|\|f'\|\|_{[a,b],p}, & \\ \qquad\text{if } f'\in L_p([a,b];X);\ p>1,\ \frac{1}{p}+\frac{1}{q}=1, & \\ \left[\frac{1}{2}+\left|s-\frac{a+b}{2}\right|\right]\|\|f'\|\|_{[a,b],1}, & \end{cases} \tag{3}$$

for any $s \in [a,b]$, *where*

$$\left\| |f'| \right\|_{[a,b],\infty} := ess \sup_{t\in[a,b]} \|f'(t)\|$$

and

$$\left\| |f'| \right\|_{[a,b],p} := \left(\int_a^b \|f'(t)\|^p \, dt \right)^{\frac{1}{p}}, \quad p \geq 1.$$

In this paper we establish similar results for functions f defined on segments in general linear spaces. We recall that, if $a, b \in X$, $a \neq b$ and X is a real linear space, then by the segment generated by a and b we understand the set

$$[a,b] := \{x \in X | x = (1-\lambda)\, a + \lambda b,\ \lambda \in [0,1]\}.$$

Natural applications for semi-inner products in normed linear spaces are also provided.

2 Ostrowski Type Inequalities

The following result of Ostrowski type for functions defined on linear spaces holds.

Theorem 4 *Let X be a linear space, $a, b \in X$, $a \neq b$ and $f : [a,b] \subset X \to \mathbb{R}$ be a function defined on the segment $[a,b]$ and such that the Gâteaux differential $\nabla f[(1-\cdot)\, a + \cdot b]\,(b-a)$ exists a.e. on $[0,1]$. Then for any $s \in [0,1]$ we have the inequalities:*

$$\left| f[(1-s)\, a + sb] - \int_0^1 f[(1-t)\, a + tb]\, dt \right|$$
$$\leq \begin{cases} \left[\frac{1}{4} + \left(s - \frac{1}{2} \right)^2 \right] \|\nabla f[(1-\cdot)\, a + \cdot b]\,(b-a)\|_\infty, \\ \qquad\qquad if\ \ \nabla f[(1-\cdot)\, a + \cdot b]\,(b-a) \in L_\infty[0,1]; \\ \\ \frac{1}{(q+1)^{\frac{1}{q}}} \left[s^{q+1} + (1-s)^{q+1} \right]^{\frac{1}{q}} \|\nabla f[(1-\cdot)\, a + \cdot b]\,(b-a)\|_p, \\ \quad if\ \ \nabla f[(1-\cdot)\, a + \cdot b]\,(b-a) \in L_p[0,1],\ p > 1,\ \frac{1}{p} + \frac{1}{q} = 1; \\ \\ \left[\frac{1}{2} + \left| s - \frac{1}{2} \right| \right] \|\nabla f[(1-\cdot)\, a + \cdot b]\,(b-a)\|_1, \end{cases} \tag{4}$$

where $\|\cdot\|_r$ $(r \in [1,\infty])$ are the usual Lebesgue norms on $L_r[a,b]$, i.e.,

$$\|\nabla f[(1-\cdot)\, a + \cdot b]\,(b-a)\|_\infty := ess \sup_{u\in[0,1]} |\nabla f[(1-u)\, a + ub]\,(b-a)|$$

and

$$\|\nabla f[(1-\cdot)\, a + \cdot b]\,(b-a)\|_p := \left(\int_0^1 |\nabla f[(1-u)\, a + ub]\,(b-a)|^p \, du \right)^{\frac{1}{p}}, \quad p \geq 1.$$

The constants $\frac{1}{4}$, $\frac{1}{(q+1)^{\frac{1}{q}}}$ *and* $\frac{1}{2}$ *respectively are sharp in the sense that they cannot be replaced by smaller constants.*

Proof. If we consider the auxiliary function $g:[0,1]\to\mathbb{R}$,

$$g(t):=f[(1-t)a+tb]$$

then we observe that for any $s\in(0,1)$

$$\begin{aligned} g'(s) &= \lim_{t\to s}\frac{g(t)-g(s)}{t-s}=\lim_{t\to s}\frac{f[(1-t)a+tb]-f[(1-s)a+sb]}{t-s} \\ &= \lim_{h\to 0}\frac{f[(1-s)a+sb+h(b-a)]-f[(1-s)a+sb]}{h} \\ &= \nabla f[(1-s)a+sb](b-a) \end{aligned}$$

showing that s is a point of differentiability for g iff the Gâteaux differential of f in the point $(1-s)a+sb$ along the direction $b-a$ exists.

Since we assumed that $\nabla f[(1-\cdot)a+\cdot b](b-a)$ exists a.e. on $[0,1]$, it follows that g is absolutely continuous on $[0,1]$ and by Theorem 2 we may state that

$$\begin{aligned} &\left|g(s)-\int_0^1 g(t)\,dt\right| \\ &\leq\begin{cases} \left[\frac{1}{4}+\left(s-\frac{1}{2}\right)^2\right]\|g'\|_\infty, & \text{if } g'\in L_\infty[0,1]; \\ \frac{1}{(q+1)^{\frac{1}{q}}}\left[s^{q+1}+(1-s)^{q+1}\right]^{\frac{1}{q}}\|g'\|_p, & \text{if } g'\in L_p[0,1],\ p>1,\ \frac{1}{p}+\frac{1}{q}=1; \\ \left[\frac{1}{2}+\left|s-\frac{1}{2}\right|\right]\|g'\|_1, \end{cases} \end{aligned}$$

for any $s\in[0,1]$, and the inequalities (4) are proved.

The sharpness of the constants follows by the scalar case. ■

The best inequality we can get from (4) is embodied in the following corollary.

Corollary 1 *With the assumptions of Theorem 4, we have the inequality:*

$$\begin{aligned} &\left|f\left(\frac{a+b}{2}\right)-\int_0^1 f[(1-t)a+tb]\,dt\right| \\ &\leq\begin{cases} \frac{1}{4}\|\nabla f[(1-\cdot)a+\cdot b](b-a)\|_\infty, \\ \text{if } \nabla f[(1-\cdot)a+\cdot b](b-a)\in L_\infty[0,1]; \\ \frac{1}{2(q+1)^{\frac{1}{q}}}\|\nabla f[(1-\cdot)a+\cdot b](b-a)\|_p, \\ \text{if } \nabla f[(1-\cdot)a+\cdot b](b-a)\in L_p[0,1]; \\ \frac{1}{2}\|\nabla f[(1-\cdot)a+\cdot b](b-a)\|_1. \end{cases} \end{aligned} \tag{5}$$

The constants $\frac{1}{4}$, $\frac{1}{2(q+1)^{\frac{1}{q}}}$ *and* $\frac{1}{2}$ *are sharp in the sense mentioned above.*

Remark 1 *If the function* $f:[a,b]\subset X\to\mathbb{R}$ *is convex on* $[a,b]$ *then* $\nabla f\left[(1-\cdot)\,a+\cdot b\right](b-a)$ *exists a.e. on* $[0,1]$ *and*

$$\nabla f\left[(1-s)\,a+sb\right](b-a)=\nabla_{\pm}f\left[(1-s)\,a+sb\right](b-a)$$

for a.e. $s\in[0,1]$, *where* $\nabla_{\pm}f\left[(1-s)\,a+sb\right](b-a)$ *are the Gâteaux lateral derivatives which exist in every point* $s\in(0,1)$.

If we recall the Hermite-Hadamard integral inequality for convex functions defined on the segment $[a,b]\subset X$,

$$f\left(\frac{a+b}{2}\right)\leq\int_0^1 f\left[(1-t)\,a+tb\right]dt\leq\frac{f(a)+f(b)}{2},\tag{6}$$

then the above corollary applied for convex functions $f:[a,b]\subset X\to\mathbb{R}$ provides the following counterpart inequality for the first inequality in (6):

$$\begin{aligned}0\quad&\leq\int_0^1 f\left[(1-t)\,a+tb\right]dt-f\left(\frac{a+b}{2}\right)\\&\leq\begin{cases}\dfrac{1}{4}\left\|\nabla_{\pm}f\left[(1-\cdot)\,a+\cdot b\right](b-a)\right\|_{\infty},\\\qquad\qquad\text{if }\ \nabla_{\pm}f\left[(1-\cdot)\,a+\cdot b\right](b-a)\in L_{\infty}[0,1];\\[2ex]\dfrac{1}{2(q+1)^{\frac{1}{q}}}\left\|\nabla_{\pm}f\left[(1-\cdot)\,a+\cdot b\right](b-a)\right\|_{p},\\\text{if }\ \nabla_{\pm}f\left[(1-\cdot)\,a+\cdot b\right](b-a)\in L_p[0,1],\ p>1,\ \frac{1}{p}+\frac{1}{q}=1;\\[2ex]\dfrac{1}{2}\left\|\nabla_{\pm}f\left[(1-\cdot)\,a+\cdot b\right](b-a)\right\|_{1}.\end{cases}\end{aligned}\tag{7}$$

3 Applications for Semi-Inner Products

Now, assume that $(X,\|\cdot\|)$ is a normed linear space. The function $f_0(x)=\frac{1}{2}\|x\|^2$, $x\in X$ is convex and thus the following limits exist

(iv) $\langle x,y\rangle_s:=\left(\nabla_+f_0(y)\right)(x)=\lim\limits_{t\to0+}\left[\frac{\|y+tx\|^2-\|y\|^2}{2t}\right]$;

(v) $\langle x,y\rangle_i:=\left(\nabla_-f_0(y)\right)(x)=\lim\limits_{s\to0-}\left[\frac{\|y+sx\|^2-\|y\|^2}{2s}\right]$;

for any $x,y\in X$. They are called the *lower* and *upper semi-inner* products associated to the norm $\|\cdot\|$.

For the sake of completeness, we list here some of the main properties of the mappings that will be used in the sequel assuming that $p,q\in\{s,i\}$ and $p\neq q$:

(a) $\langle x,x\rangle_p=\|x\|^2$ for all $x\in X$;

(aa) $\langle \alpha x, \beta y\rangle_p = \alpha\beta \langle x, y\rangle_p$ of $\alpha, \beta \geq 0$ and $x, y \in X$;

(aaa) $\left|\langle x, y\rangle_p\right| \leq \|x\| \|y\|$ for all $x, y \in X$;

(av) $\langle \alpha x + y, x\rangle_p = \alpha \langle x, x\rangle_p + \langle y, x\rangle_p$ if $x, y \in X$ and $\alpha \in \mathbb{R}$;

(v) $\langle -x, y\rangle_p = -\langle x, y\rangle_q$ for all $x, y \in X$;

(va) $\langle x + y, z\rangle_p \leq \|x\| \|z\| + \langle y, z\rangle_p$ for all $x, y, z \in X$;

(vaa) The mapping $\langle \cdot, \cdot\rangle_p$ is continuous and subadditive (superadditive) in the first variable for $p = 0$ (or $p = i$);

(vaaa) The normed linear space $(X, \|\cdot\|)$ is smooth at the point $x_0 \in X \setminus \{0\}$ if and only if $\langle y, x_0\rangle_s = \langle y, x_0\rangle_i$ for all $y \in X$; in general $\langle y, x\rangle_i \leq \langle y, x\rangle_s$ for all $x, y \in X$;

(ax) If the norm $\|\cdot\|$ is induced by an inner product $\langle \cdot, \cdot\rangle$, then $\langle y, x\rangle_i = \langle y, x\rangle = \langle y, x\rangle_s$ for all $x, y \in X$.

Now, if we apply Theorem 4 for the function $f_0 : X \to \mathbb{R}$, $f_0(x) = \frac{1}{2}\|x\|^2$, then we may state the following proposition:

Proposition 1 *For any $a, b \in X$ and $s \in [0, 1]$, one has the inequality:*

$$\left|\|(1-s)a + sb\|^2 - \int_0^1 \|(1-t)a + tb\|^2 dt\right| \tag{8}$$

$$\leq 2\times \begin{cases} \left[\frac{1}{4} + \left(s - \frac{1}{2}\right)^2\right] \sup\limits_{u\in[0,1]} \left|\langle b-a, (1-u)a + ub\rangle_{s(i)}\right|, \\ \\ \frac{1}{(q+1)^{\frac{1}{q}}} \left[s^{q+1} + (1-s)^{q+1}\right]^{\frac{1}{q}} \left(\int_0^1 \left|\langle b-a, (1-u)a + ub\rangle_{s(i)}\right|^p du\right)^{\frac{1}{p}}, \\ \qquad\qquad p > 1,\ \frac{1}{p} + \frac{1}{q} = 1; \\ \\ \left[\frac{1}{2} + \left|s - \frac{1}{2}\right|\right] \int_0^1 \left|\langle b-a, (1-u)a + ub\rangle_{s(i)}\right| du \end{cases} \tag{9}$$

and, in particular,

$$0 \leq \int_0^1 \|(1-t)\,a+tb\|^2\,dt - \left\|\frac{a+b}{2}\right\|^2 \leq \begin{cases} \frac{1}{2}\sup\limits_{u\in[0,1]}\left|\langle b-a,(1-u)\,a+ub\rangle_{s(i)}\right|, \\ \frac{1}{(q+1)^{\frac{1}{q}}}\left(\int_0^1\left|\langle b-a,(1-u)\,a+ub\rangle_{s(i)}\right|^p du\right)^{\frac{1}{p}}, \\ \qquad\qquad p>1,\ \frac{1}{p}+\frac{1}{q}=1; \\ \int_0^1\left|\langle b-a,(1-u)\,a+ub\rangle_{s(i)}\right| du. \end{cases} \tag{10}$$

Remark 2 *Since*

$$\left|\langle b-a,(1-u)\,a+ub\rangle_{s(i)}\right| \leq \|b-a\|\,\|(1-u)\,a+ub\|$$

then

$$\sup_{u\in[0,1]}\left|\langle b-a,(1-u)\,a+ub\rangle_{s(i)}\right| \leq \|b-a\|\sup_{u\in[0,1]}\|(1-u)\,a+ub\| = \|b-a\|\max\{\|a\|,\|b\|\},$$

$$\left(\int_0^1\left|\langle b-a,(1-u)\,a+ub\rangle_{s(i)}\right|^p du\right)^{\frac{1}{p}} \leq \|b-a\|\left(\int_0^1\|(1-u)\,a+ub\|^p\,du\right)^{\frac{1}{p}}$$

and

$$\int_0^1\left|\langle b-a,(1-u)\,a+ub\rangle_{s(i)}\right| du \leq \|b-a\|\int_0^1\|(1-u)\,a+ub\|\,du$$

and by (8) we may deduce the inequalities

$$\left|\|(1-s)\,a+sb\|^2 - \int_0^1\|(1-t)\,a+tb\|^2\,dt\right| \tag{11}$$

$$\leq 2\,\|b-a\|\times\begin{cases} \max\{\|a\|,\|b\|\}\left[\frac{1}{4}+\left(s-\frac{1}{2}\right)^2\right], \\ \left(\int_0^1\|(1-u)\,a+ub\|^p\,du\right)^{\frac{1}{p}}\left[\frac{1}{(q+1)^{\frac{1}{q}}}\left[s^{q+1}+(1-s)^{q+1}\right]^{\frac{1}{q}}\right], \\ \qquad\qquad p>1,\ \frac{1}{p}+\frac{1}{q}=1; \\ \int_0^1\|(1-u)\,a+ub\|\,du\cdot\left[\frac{1}{2}+\left|s-\frac{1}{2}\right|\right] \end{cases} \tag{12}$$

and, in particular,

$$0 \leq \int_0^1 \|(1-t)a+tb\|^2 dt - \left\|\frac{a+b}{2}\right\|^2$$
$$\leq \|b-a\| \times \begin{cases} \frac{1}{2}\max\{\|a\|,\|b\|\}, \\ \\ \frac{1}{(q+1)^{\frac{1}{q}}}\left(\int_0^1 \|(1-u)a+ub\|^p du\right)^{\frac{1}{p}}, \\ \qquad\qquad p>1,\ \frac{1}{p}+\frac{1}{q}=1; \\ \\ \int_0^1 \|(1-u)a+ub\| du. \end{cases} \tag{13}$$

Now, if we choose the function $f_0 : X \to \mathbb{R}$, $f_0(x) = \|x\|$, then we may state the following proposition as well.

Proposition 2 *For any a, b linearly independent vectors in X and $s \in [0,1]$, one has the inequalities:*

$$\left|\|(1-s)a+sb\| - \int_0^1 \|(1-t)a+tb\| dt\right|$$
$$\leq \begin{cases} \left[\frac{1}{4}+\left(s-\frac{1}{2}\right)^2\right]\sup\limits_{u\in[0,1]}\left|\frac{\langle b-a,(1-u)a+ub\rangle_{s(i)}}{\|(1-u)a+ub\|}\right|, \\ \\ \frac{1}{(q+1)^{\frac{1}{q}}}\left[s^{q+1}+(1-s)^{q+1}\right]^{\frac{1}{q}}\left(\int_0^1\left|\frac{\langle b-a,(1-u)a+ub\rangle_{s(i)}}{\|(1-u)a+ub\|}\right|^p du\right)^{\frac{1}{p}}, \\ \qquad\qquad p>1,\ \frac{1}{p}+\frac{1}{q}=1; \\ \\ \left[\frac{1}{2}+\left|s-\frac{1}{2}\right|\right]\int_0^1\left|\frac{\langle b-a,(1-u)a+ub\rangle_{s(i)}}{\|(1-u)a+ub\|}\right| du \end{cases} \tag{14}$$
$$\leq \begin{cases} \left[\frac{1}{4}+\left(s-\frac{1}{2}\right)^2\right]\|b-a\|, \\ \\ \frac{1}{(q+1)^{\frac{1}{q}}}\left[s^{q+1}+(1-s)^{q+1}\right]^{\frac{1}{q}}\|b-a\|, \\ \qquad\qquad p>1,\ \frac{1}{p}+\frac{1}{q}=1; \\ \\ \left[\frac{1}{2}+\left|s-\frac{1}{2}\right|\right]\|b-a\|, \end{cases}$$

and, in particular,

$$0 \leq \int_0^1 \|(1-t)\,a+tb\|\,dt - \left\|\frac{a+b}{2}\right\| \leq \begin{cases} \dfrac{1}{4}\sup\limits_{u\in[0,1]}\left|\dfrac{\langle b-a,(1-u)\,a+ub\rangle_{s(i)}}{\|(1-u)\,a+ub\|}\right|, \\ \dfrac{1}{2\,(q+1)^{\frac{1}{q}}}\left(\displaystyle\int_0^1\left|\dfrac{\langle b-a,(1-u)\,a+ub\rangle_{s(i)}}{\|(1-u)\,a+ub\|}\right|^p du\right)^{\frac{1}{p}}, \\ \qquad\qquad p>1,\ \frac{1}{p}+\frac{1}{q}=1; \\ \dfrac{1}{2}\displaystyle\int_0^1\left|\dfrac{\langle b-a,(1-u)\,a+ub\rangle_{s(i)}}{\|(1-u)\,a+ub\|}\right| du. \end{cases} \tag{15}$$

From the first inequality in (15), we get the simple and nice result :

$$0 \leq \int_0^1 \|(1-t)\,a+tb\|\,dt - \left\|\frac{a+b}{2}\right\| \leq \frac{1}{4}\,\|b-a\|\,, \tag{16}$$

holding in any normed linear space X and for any $a,b\in X$.

The constant $\frac{1}{4}$ is sharp.

Indeed, if we assume that (16) holds with a constant $C>0$, i.e.,

$$0 \leq \int_0^1 \|(1-t)\,a+tb\|\,dt - \left\|\frac{a+b}{2}\right\| \leq C\,\|b-a\| \tag{17}$$

and choose $X=\mathbb{R}$, $\|\cdot\|=|\cdot|$, $a=-1$, $b=1$, then we get

$$\int_0^1 \|(1-t)\,a+tb\|\,dt = \int_0^1 |2t-1|\,dt = \frac{1}{2}$$
$$\left\|\frac{a+b}{2}\right\| = 0, \quad \|b-a\| = 2$$

and (17) becomes

$$\frac{1}{2} \leq 2C,$$

showing that $C\geq\frac{1}{4}$.

4 Applications for Functions of Several Variables

Now, let $\Omega\subset\mathbb{R}^n$ be an open convex set in $\mathbb{R}^n$.

If $F:\Omega\to\mathbb{R}$ is a differentiable function on Ω, then, obviously, for any $\bar{c}\in\Omega$ we have

$$\nabla F\,(\bar{c})\,(\bar{y}) = \sum_{i=1}^n \frac{\partial F\,(\bar{c})}{\partial x_i}\cdot y_i, \quad \bar{y}\in\mathbb{R}^n,$$

where $\frac{\partial F}{\partial x_i}$ are the partial derivatives of F with respect to the variable x_i $(i=1,\ldots,n)$.

Using Theorem 4, we may state the following inequality:

$$\left|F\left[(1-s)\,\bar{a}+s\bar{b}\right]-\int_0^1 F\left[(1-t)\,\bar{a}+t\bar{b}\right]dt\right| \tag{18}$$

$$\leq\begin{cases}\left[\frac{1}{4}+\left(s-\frac{1}{2}\right)^2\right]\sup\limits_{u\in[0,1]}\left|\sum\limits_{i=1}^{n}\frac{\partial F\left((1-u)\,\bar{a}+u\bar{b}\right)}{\partial x_i}\left(b_i-a_i\right)\right| \\ \frac{1}{(q+1)^{\frac{1}{q}}}\left[s^{q+1}+(1-s)^{q+1}\right]^{\frac{1}{q}}\left(\int_0^1\left|\sum\limits_{i=1}^{n}\frac{\partial F\left((1-u)\,\bar{a}+u\bar{b}\right)}{\partial x_i}\cdot\left(b_i-a_i\right)\right|^p du\right)^{\frac{1}{p}} \\ \left[\frac{1}{2}+\left|s-\frac{1}{2}\right|\right]\int_0^1\left|\sum\limits_{i=1}^{n}\frac{\partial F\left((1-u)\,\bar{a}+u\bar{b}\right)}{\partial x_i}\cdot\left(b_i-a_i\right)\right|du.\end{cases}$$

Now, observe that

$$\left|\sum_{i=1}^{n}\frac{\partial F\left((1-u)\,\bar{a}+u\bar{b}\right)}{\partial x_i}\cdot\left(b_i-a_i\right)\right|$$
$$\leq\begin{cases}\left\|\frac{\partial F\left((1-u)\,\bar{a}+u\bar{b}\right)}{\partial\bar{x}}\right\|_\infty\left|\left|\left|b-a\right|\right|\right|_1 \\ \left\|\frac{\partial F\left((1-u)\,\bar{a}+u\bar{b}\right)}{\partial\bar{x}}\right\|_\alpha\left|\left|\left|b-a\right|\right|\right|_\beta \\ \left\|\frac{\partial F\left((1-u)\,\bar{a}+u\bar{b}\right)}{\partial\bar{x}}\right\|_1\left|\left|\left|b-a\right|\right|\right|_\infty\end{cases},$$

where $\left|\left|\left|\cdot\right|\right|\right|_\ell$ $(\ell\in[1,\infty])$ are the $\ell-$norms in $\mathbb{R}^n$, i.e.,

$$\left|\left|\left|\bar{v}\right|\right|\right|_1 := \sum_{i=1}^{n}\left|v_i\right|,$$
$$\left|\left|\left|\bar{v}\right|\right|\right|_\alpha := \left(\sum_{i=1}^{n}\left|v_i\right|^\alpha\right)^{\frac{1}{\alpha}},\ \ \alpha>1$$

and

$$\left|\left|\left|\bar{v}\right|\right|\right|_\infty := \max_{i=\overline{1,n}}\left|v_i\right|,$$

whereas

$$\frac{\partial F\left((1-u)\,\bar{a}+u\bar{b}\right)}{\partial\bar{x}}=\left(\frac{\partial F\left((1-u)\,\bar{a}+u\bar{b}\right)}{\partial x_1},\cdots,\frac{\partial F\left((1-u)\,\bar{a}+u\bar{b}\right)}{\partial x_n}\right).$$

Consequently, we get

$$\sup_{u\in[0,1]}\left|\sum_{i=1}^{n}\frac{\partial F\left(\left(1-u\right)\bar{a}+u\bar{b}\right)}{\partial x_i}\cdot\left(b_i-a_i\right)\right| \tag{19}$$

$$\leq\begin{cases}\left|\left|\left|b-a\right|\right|\right|_1\sup\limits_{u\in[0,1]}\left\|\left\|\dfrac{\partial F\left(\left(1-u\right)\bar{a}+u\bar{b}\right)}{\partial\bar{x}}\right\|\right\|_\infty\\ \left|\left|\left|b-a\right|\right|\right|_\beta\sup\limits_{u\in[0,1]}\left\|\left\|\dfrac{\partial F\left(\left(1-u\right)\bar{a}+u\bar{b}\right)}{\partial\bar{x}}\right\|\right\|_\alpha, \quad \frac{1}{\alpha}+\frac{1}{\beta}=1,\ \alpha>1\ ,\\ \left|\left|\left|b-a\right|\right|\right|_\infty\sup\limits_{u\in[0,1]}\left\|\left\|\dfrac{\partial F\left(\left(1-u\right)\bar{a}+u\bar{b}\right)}{\partial\bar{x}}\right\|\right\|_1\end{cases}$$

and

$$\left(\int_0^1\left|\sum_{i=1}^{n}\frac{\partial F\left(\left(1-u\right)\bar{a}+u\bar{b}\right)}{\partial x_i}\cdot\left(b_i-a_i\right)\right|^p du\right)^{\frac{1}{p}} \tag{20}$$

$$\leq\begin{cases}\left|\left|\left|b-a\right|\right|\right|_1\left(\int_0^1\left\|\left\|\dfrac{\partial F\left(\left(1-u\right)\bar{a}+u\bar{b}\right)}{\partial\bar{x}}\right\|\right\|_\infty^p du\right)^{\frac{1}{p}}\\ \left|\left|\left|b-a\right|\right|\right|_\beta\left(\int_0^1\left\|\left\|\dfrac{\partial F\left(\left(1-u\right)\bar{a}+u\bar{b}\right)}{\partial\bar{x}}\right\|\right\|_\alpha^p du\right)^{\frac{1}{p}},\\ \left|\left|\left|b-a\right|\right|\right|_\infty\left(\int_0^1\left\|\left\|\dfrac{\partial F\left(\left(1-u\right)\bar{a}+u\bar{b}\right)}{\partial\bar{x}}\right\|\right\|_1^p du\right)^{\frac{1}{p}}\end{cases}$$

for $p\geq 1$.

In practical applications we usually assume that

$$\left\|\frac{\partial F}{\partial x_i}\right\|_{\Omega,\infty}:=\sup_{\bar{u}\in\bar{\Omega}}\left|\frac{\partial F\left(\bar{u}\right)}{\partial x_i}\right|<\infty.$$

Then, by the first inequality in (18) we deduce:

$$\begin{aligned}&\left|F\left[\left(1-s\right)\bar{a}+s\bar{b}\right]-\int_0^1 F\left[\left(1-t\right)\bar{a}+t\bar{b}\right]dt\right|\\ &\qquad\leq\left[\frac{1}{4}+\left(s-\frac{1}{2}\right)^2\right]\sum_{i=1}^{n}\left\|\frac{\partial F}{\partial x_i}\right\|_{\Omega,\infty}\left|b_i-a_i\right|,\end{aligned} \tag{21}$$

for any $s\in[0,1]$.

In particular, we have the following mid-point inequality

$$\left|F\left(\frac{\bar{a}+\bar{b}}{2}\right)-\int_0^1 F\left[(1-t)\,\bar{a}+t\bar{b}\right]dt\right| \leq \frac{1}{4}\sum_{i=1}^{n}\left\|\frac{\partial F}{\partial x_i}\right\|_{\Omega,\infty}|b_i-a_i|\,. \qquad (22)$$

References

[1] G.A. ANASTASSIOU, Ostrowski type inequalities, *Proc. of Amer. Math. Soc.,* **123**(12) (1999), 3775-3781.

[2] N.S. BARNETT, C. BUŞE, P. CERONE and S.S. DRAGOMIR, Ostrowski's inequality for vector-valued functions and applications, *Comp. & Math. Appl.,* (accepted).

[3] P. CERONE and S.S. DRAGOMIR, Midpoint-type rules from an inequalities point of view, *Handbook of Analytic-Computational Methods in Applied Mathematics,* Editor: G. Anastassiou, CRC Press, New York (2000), 135-200.

[4] P. CERONE and S.S. DRAGOMIR, Three point quadrature rules involving, at most, a first derivative, (submitted) (ON LINE: `http://rgmia.vu.edu.au/v2n4.html`).

[5] P. CERONE and S.S. DRAGOMIR, Trapezoidal-type rules from an inequalities point of view, *Handbook of Analytic-Computational Methods in Applied Mathematics,* Editor: G. Anastassiou, CRC Press, New York (2000), 65-134.

[6] S.S. DRAGOMIR, A generalization of Ostrowski integral inequality for mappings whose derivatives belong to $L_\infty[a,b]$ and applications in numerical integration, *J. KSIAM,* **5**(2) (2001), 117-136.

[7] S.S. DRAGOMIR, A generalization of Ostrowski integral inequality for mappings whose derivatives belong to $L_1[a,b]$ and applications in numerical integration, *J. of Computational Analysis and Applications,* **3**(4) (2001), 343-360.

[8] S.S. DRAGOMIR, A generalization of Ostrowski integral inequality for mappings whose derivatives belong to $L_p[a,b]$, $1<p<\infty$, and applications in numerical integration, *J. Math. Anal. Appl.,* **225** (2001), 605-626.

[9] S.S. DRAGOMIR, On Simpson's quadrature formula for differentiable mappings whose derivatives belong to L_p-spaces and applications, *J. KSIAM*, **2** (1998), 57-65.

[10] S.S. DRAGOMIR, On Simpson's quadrature formula for mappings of bounded variation and applications, *Tamkang J. of Math.,* **30** (1)(1999), 53-58.

[11] S.S. DRAGOMIR, On Simpson's quadrature formula for Lipschitzian mappings and applications, *Soochow J. of Math.,* **25** (2)(1999), 175-180.

[12] S.S. DRAGOMIR, On the Ostrowski's integral inequality for mappings with bounded variation and applications, *Math. Ineq. & Appl.,* **4**(1) (2001), 59-66. (ON LINE: `http://rgmia.vu.edu.au/v2n1.html`).

[13] S.S. DRAGOMIR, Ostrowski inequality for monotonous mappings and applications, *J. KSIAM*, **3**(1) (1999), 127-135.

[14] S.S. DRAGOMIR, The Ostrowski's integral inequality for Lipschitzian mappings and applications, *Comp. and Math. with Appl.,* **38** (1999), 33-37.

[15] S.S. DRAGOMIR, The Ostrowski's integral inequality for mappings of bounded variation, *Bull. Austral. Math. Soc.,* **60** (1999), 495-508.

[16] S.S. DRAGOMIR, P. CERONE, J. ROUMELIOTIS and S. WANG, A weighted version of Ostrowski inequality for mappings of Hölder type and applications in numerical analysis, *Bull. Math. Soc. Sc. Math. Roumanie,* **42**(90)(4) (1992), 301-304.

[17] S.S. DRAGOMIR, J. PEČARIĆ and S. WANG, The unified treatment of trapezoid, Simpson and Ostrowski type inequalities for monotonic mappings and applications, *Math. and Comp. Modelling,* **31** (2000), 61-70.

[18] S.S. DRAGOMIR and Th. M. RASSIAS (Ed.), *Ostrowski Type Inequalities and Applications in Numerical Integration,* Kluwer Academic Publishers, 2002.

[19] S.S. DRAGOMIR and S. WANG, A new inequality of Ostrowski's type in L_1-norm and applications to some special means and to some numerical quadrature rules, *Tamkang J. of Math.*, **28** (1997), 239-244.

[20] S.S. DRAGOMIR and S. WANG, A new inequality of Ostrowski's type in L_p-norm, *Indian Journal of Mathematics,* **40**(3) (1998), 299-304.

[21] S.S. DRAGOMIR and S. WANG, An inequality of Ostrowski-Grüss' type and its applications to the estimation of error bounds for some special means and for some numerical quadrature rules, *Computers Math. Applic.,* **33** (1997), 15-20.

[22] S.S. DRAGOMIR and S. WANG, Applications of Ostrowski's inequality to the estimation of error bounds for some special means and some numerical quadrature rules, *Appl. Math. Lett.*, **11** (1998), 105-109.

[23] A.M. FINK, Bounds on the derivation of a function from its averages, *Czech. Math. J.,* **42** (1992), 289-310.

[24] D.S. MITRINOVIĆ, J.E. PEČARIĆ and A.M. FINK, *Inequalities Involving Functions and Their Integrals and Derivatives,* Kluwer Academic Publishers, Dordrecht, 1994.

JOURNAL OF CONCRETE AND APPLICABLE MATHEMATICS,VOL.3,NO.1,105-114,2005,COPYRIGHT 2005 EUDOXUS PRESS,LLC

Multivariate Steinberg Theorem

Daniel Boichu and Vincent Robin
Université de Technologie de Compiègne
GI/LMAC
BP 20529 - 60205 COMPIEGNE Cedex, FRANCE
Daniel.Boichu@utc.fr
Vincent.Robin@utc.fr

Abstract

We prove the following theorem, already proved by S. Steinberg [6] in one variable. Let $d \geq 1$ be an integer and $\Omega \subset \mathbb{C}^d$ be a connected open set. Suppose that the mapping

$$\begin{array}{rcl} T : \Omega & \longrightarrow & \mathcal{B}(X) \\ z & \longmapsto & T(z) \end{array}$$

is holomorphic in Ω, and that $T(z)$ is compact for all $z \in \Omega$. Then we have the following alternative:

(i) either for all $z \in \Omega$, $e - T(z)$ is not invertible;

(ii) either $z \mapsto [e - T(z)]^{-1}$ is meromorphic in Ω.

When $dim X < \infty$, it is a consequence of Cramer's rule. When $dim X = \infty$, the proof is based on spectral theory and complex analysis.

Keywords
holomorphy and meromorphy for multivariate operator-valued functions, spectral theory, meromorphy of $z \mapsto (e - T(z))^{-1}$ when $T(z)$ is compact.

AMS codes
32A30, 47A13, 35P99

1 Introduction

Throughout this article, X denotes a complex Banach space and $\mathcal{B}(X)$ is the set of bounded linear operators on X. $\mathcal{B}(X)$ is a Banach algebra (with operator composition as multiplication) and we denote its unit by e (as Rudin, see [5, p.227-230] for details).

We aim to prove the following

Theorem 1 *Let $d \geq 1$ be an integer and $\Omega \subset \mathbb{C}^d$ be a connected open set. Suppose that the mapping*

$$\begin{aligned} T : \Omega &\longrightarrow \mathcal{B}(X) \\ z &\longmapsto T(z) \end{aligned}$$

is holomorphic in Ω, and that $T(z)$ is compact for all $z \in \Omega$. Then we have the following alternative:

(i) either for all $z \in \Omega$, $e - T(z)$ is not invertible;

(ii) either $z \mapsto [e - T(z)]^{-1}$ is meromorphic in Ω.

Steinberg [6] proved this theorem in the case $d = 1$.

Motivation. The typical situation where the Steinberg theorem should be used is the following. Suppose that a linear partial differential equation (PDE) problem can be stated as :

$$\begin{cases} \forall z \in \Omega, \text{ for } b(z) \in Y, \text{ find } u(z) \in X \text{ such that} \\ A(z)u(z) = b(z) \end{cases}$$

where X and Y are Banach spaces, $z \mapsto A(z)$ is holomorphic in Ω and $A(z) \in \mathcal{B}(X, Y)$ is an isomorphism minus a compact perturbation for all $z \in \Omega$:

$$A(z) = \underbrace{\Theta_0}_{isomorphism} - \underbrace{\Theta(z)}_{compact} = \Theta_0[e - \underbrace{\Theta_0^{-1}\Theta(z)}_{compact}] \tag{1}$$

Thus $T(z) = \Theta_0^{-1}\Theta(z) \in \mathcal{B}(X)$ is compact for all $z \in \Omega$ and holomorphic in Ω.

For example, using this technique, the electric field scattered by a perfectly conducting object is shown to be meromorphic with respect to the frequency [4]. In order to deal with problems depending on more than one parameter, we have to generalize this Steinberg theorem to compact operators $T(z)$ with $z \in \mathbb{C}^d$, where $d \geq 1$ denotes the number of parameters of the problem. Note also that meromorphy is a necessary condition for convergence of Padé approximants (see [1]).

The outline is as follow. In section 2, we recall some elements of spectral theory. In section 3, we define holomorphic multivariate X-valued function. In section 4, we state precisely the notion of meromorphy for operator-valued functions in $\mathbb{C}^d$. In particular we emphasize that in finite dimension, the Steinberg theorem reduces to Cramer's rule. In section 5, we prove the Steinberg theorem for several variables.

2 Some basic facts about operators

Our proof of theorem 1 will make wide use of symbolic calculus and spectral properties of compact operators. We recall here, without proof, only what is needed to understand the proof of theorem 1.

Lemma 1 *[5, p.235]* $G = \{T \in \mathcal{B}(X) : \ T \text{ invertible}\}$ *is an open subset of* $\mathcal{B}(X)$ *and the mapping*

$$\begin{array}{rccc} \mathcal{I} : G & \longrightarrow & G \\ T & \mapsto & T^{-1} \end{array}$$

is a diffeomorphism.

The spectrum of $T \in \mathcal{B}(X)$ is

$$\sigma(T) = \{\lambda \in \mathbb{C} \ : \quad \lambda e - T \ \text{ not invertible}\}$$

$\sigma(T) \neq \emptyset$ and $\sigma(T)$ is compact (closed and bounded) in $\mathbb{C}$ (see Rudin [5, p.235]). Moreover $\sigma(T)$ changes continuously with T, precisely :

Lemma 2 *[5, p.239] Suppose* $T \in \mathcal{B}(X)$, O *is an open set in* $\mathbb{C}$, *and* $\sigma(T) \subset O$. *Then there exists* $\epsilon > 0$ *such that* $\sigma(T+S) \subset O$ *for every* $S \in \mathcal{B}(X)$ *with* $|S| < \epsilon$.

Symbolic calculus. (see Rudin [5, p.240-248] for details) If $P(z) = a_0 + a_1 z + \cdots + a_n z^n$ is a polynomial, we can define $P(T)$ for every $T \in \mathcal{B}(X)$ by the formula

$$P(T) = a_0 e + a_1 T + \cdots + a_n T^n \in \mathcal{B}(X) \tag{2}$$

Moreover, we can do the same extension for any $f \in \mathcal{O}(O)$, the algebra of complex holmorphic function on the open set $O \subset \mathbb{C}$, using a Cauchy formula. Precisely, we define a mapping $f \in \mathcal{O}(O) \mapsto \tilde{f} \in A_O = \{T \in \mathcal{B}(X) : \ \sigma(T) \subset O\}$ by the Cauchy formula :

$$\tilde{f}(T) = \frac{1}{2i\pi} \int_\Gamma f(\lambda)(\lambda e - T)^{-1} d\lambda \tag{3}$$

where Γ is any contour that surrounds $\sigma(T)$ in O. If f is polynomial, the two formula (2) and (3) coincide. In particular, for the holomorphic functions in O defined respectively by $\mathbf{1}_O : z \mapsto 1$ and $\lambda : z \mapsto z$, we have for all $T \in A_O$,

$$\widetilde{\mathbf{1}_O}(T) \quad = \quad \frac{1}{2i\pi} \int_\Gamma (\lambda e - T)^{-1} d\lambda = e \tag{4}$$

$$\widetilde{\lambda}(T) \quad = \quad \frac{1}{2i\pi} \int_\Gamma \lambda(\lambda e - T)^{-1} d\lambda = T \tag{5}$$

This mapping $f \mapsto \tilde{f}$ is an algebra isomorphism [5, p.244], in particular $\widetilde{fg} = \tilde{f}\tilde{g}$. Moreover, we have the important

Lemma 3 (spectral mapping theorem) *[5, p.244] Suppose* $T \in A_O$ *and* $f \in \mathcal{O}(O)$. *Then*

$$\sigma(\tilde{f}(T)) = f(\sigma(T))$$

It is worth noticing that all the facts mentioned above remain true when $\mathcal{B}(X)$ is replaced by any Banach algebra (see [5, p.227-259]). However, the compact elements of $\mathcal{B}(X)$ have additional crucial properties. Recall that $T \in \mathcal{B}(X)$ is compact if the closure of $T(U)$, U the open unit ball of X, is compact (in X).

Lemma 4 *[5, p.103] If $T \in \mathcal{B}(X)$ is compact, $\sigma(T)$ is at most countable, and has at most one limit point, namely,* 0.

Lemma 5 *[5, p.99] If $T \in \mathcal{B}(X)$ is compact, and $\lambda \neq 0$, then the nullspace $\mathcal{N}(\lambda e - T)$ is finite dimensional.*

3 Holomorphy for operator-valued functions

In the sequel, Ω denotes an open subset of $\mathbb{C}^d$.

Definition 1 *Let W be a complex topological vector space. $f : \Omega \longrightarrow W$ is said to be holomorphic in Ω if it is holomorphic with respect to each complex variable z_i, $i = 1, \ldots, d$, that is the complex partial derivatives*

$$\frac{\partial f}{\partial z_i}(z) = \lim_{t_i \to z_i} \frac{f(z_1, \ldots, z_{i-1}, t_i, z_{i+1}, \ldots, z_d) - f(z_1, \ldots, z_d)}{t_i - z_i} \quad , i = 1, \ldots, d$$

exist (in the topology of W) for every $z = (z_1, \ldots, z_d) \in \Omega$.

Let $\mathcal{O}(\Omega, W)$ the set of holomorphic W-valued function on Ω. In the sequel, we consider only the case where W is a complex Banach space (namely $\mathbb{C}$, X, or $\mathcal{B}(X)$). We simply write $\mathcal{O}(\Omega)$ for $\mathcal{O}(\Omega, \mathbb{C})$.

The following lemmas are very useful for proving holomorphy.

Lemma 6 *Suppose that the mapping*

$$\begin{array}{rcl} T : \Omega & \longrightarrow & \mathcal{B}(X) \\ z & \mapsto & T(z) \end{array}$$

is holomorphic and that $T(z)$ is invertible for all $z \in \Omega$. Then the mapping $z \mapsto T(z)^{-1}$ is holomorphic in Ω.

Proof. $T(z)^{-1} = \mathcal{I}(T(z))$. Use the chain rule and lemma 1. □

Lemma 7 *If the $\mathcal{B}(X)$-valued mappings $z \mapsto T(z)$ and $z \mapsto R(z)$ are holomorphic in Ω, then the mapping $z \mapsto T(z)R(z)$ is holomorphic in Ω.*

Proof. Use the chain rule, after noting that the mapping $(S, T) \mapsto ST$ from $\mathcal{B}(X) \times \mathcal{B}(X)$ to $\mathcal{B}(X)$ is bilinear continuous, hence differentiable. Or just proceed by direct calculus, using definition 1 above. □

Lemma 8 *Let U be a measurable subset of $\mathbb{R}^n$. Suppose that the mapping $T : (z, \lambda) \mapsto T(z, \lambda)$ from $\Omega \times U$ to $\mathcal{B}(X)$ is such that :*

- *$z \mapsto T(z, \lambda)$ is holomorphic almost everywhere in U;*
- *$t \mapsto T(z, \lambda)$ is measurable for every $z \in \Omega$;*

- *For each $z \in \Omega$, there exists a neighborhood V of z and a positive function $\theta : U \longrightarrow \mathbb{R}$ such that, almost evrywhere in U,*

$$\forall t \in V, \quad |T(t,\lambda)| \leq \theta(\lambda)$$

Then the mapping $z \mapsto G(z) = \int_U T(z,\lambda)d\lambda$ is holomorphic in Ω.

Proof. Let $z \in \Omega$ and V be a neighborhood of z as in the hypothesis of the lemma. By Lebesgue theorem, $G(z)$ is continuous in V. Hence we can use the Morera caracterization of holomorphic function, for each of the d complex variables. The Fubini theorem gives the result. Indeed, let Q be a triangle in the z_1-plane of V.

$$\int_Q G(z)dz_1 = \int_Q \int_U T(z,\lambda)d\lambda dz_1 = \int_U \underbrace{\left(\int_Q T(z,\lambda)dz_1\right)}_{=0} d\lambda = 0$$

□

4 Meromorphy for operator valued functions

We may define a meromorphic function $f : \Omega \to X$ as a function which can be written as $f = \frac{h}{\phi}$, with $h : \Omega \to X$ and $\phi : \Omega \to \mathbb{C}$ holomorphic. However such a global definition is rather impracticable and we use a local definition of meromorphy, adapted from Range [3, p. 233].

Definition 2 *$f : \Omega \to X$ is meromorphic in Ω if for every $a \in \Omega$ there are a connected neighborhood $U \in \Omega$ and holomorphic function $g \in \mathcal{O}(U)$ with $g \not\equiv 0$, such that gf can be holomorphically continued into U.*

Remark 1 *The main problem with a meromorphic function f in Ω is that f is not well defined as a function : $f(z)$ is not defined when $g(z) = 0$.*

Remark 2 *Connectedness of U is crucial in order to avoid that $g \equiv 0$ in an open subset of U. In other words, the germ of holomorphic function defined by g at each point $t \in U$ can not be null (see Range [3, p.233] for a definition in term of germs). However, if U happened to be not connected, we can always choose a smaller neighborhood which is connected (a polydisc for example).*

Remark 3 *In the particular case $d = 1$, every meromorphic function $f : \Omega \longrightarrow X$ where Ω is an open set of $\mathbb{C}$, can be written as $f = \frac{h}{\phi}$ with $h : \Omega \longrightarrow X$ and $\phi : \Omega \longrightarrow \mathbb{C}$ holomorphic.*
For $d \geq 2$, this result is no more true in general, but it still holds if Ω is a Stein manifold [3, p. 249]. In particular, it holds if Ω is a polydisk $B(z,R) = \left\{t \in \mathbb{C}^d; \quad |t_i - z_i| < R_i, \forall i = 1,\ldots,d\right\}$, and this is all we need to apply extrapolation by Padé approximants [2].

Obviously, $f \in \mathcal{O}(\Omega, X)$ is meromorphic in Ω and the sum of two meromorphic function is meromorphic. As a more interesting example, we show that theorem 1 when $dim X < \infty$ is nothing else than the Cramer's rule.

Lemma 9 *Let X be a complex finite dimensional vector space and Ω be a connected open subset of $\mathbb{C}$. Suppose that*

$$\begin{array}{rcl} T : \Omega & \longrightarrow & \mathcal{B}(X) \\ z & \mapsto & T(z) \end{array}$$

is holomorphic. Then we have the following alternative :

(i) either $T(z)$ is nowhere invertible ($\Leftrightarrow det(T(z)) \equiv 0$ in Ω);

(ii) either $T(z)^{-1}$ is meromorphic in Ω.

Proof. Since X is finite dimensional, we choose a base for X and identify $T(z)$ as a matrix with holomophic coefficients. Suppose that (i) is false. Let $z \in \Omega$, therefore Ω is a connected neighborhood of z. The Cramer's rule

$$T(t)^{-1} = \frac{1}{det(T(t))} T(t)^{com} \qquad if \quad det(T(t)) \neq 0$$

shows that $det(T(t))T(t)^{-1}$ can be continued into a holomorphic function in Ω, namely $T(t)^{com}$. □

5 Proof of theorem 1

Let X be a Banach space and $\Omega \subset \mathbb{C}^d$ be an open set. Let

$$\begin{array}{rcl} T : \Omega & \longrightarrow & \mathcal{B}(X) \\ z & \longmapsto & T(z) \end{array}$$

be holomorphic in Ω, and suppose that $T(z)$ is compact for all $z \in \Omega$.

If $e - T(z)$ is nowhere invertible, there is nothing to prove.

Suppose exists $z \in \Omega$ for which $e - T(z)$ is invertible. We have to show that $[e - T(z)]^{-1}$ is meromorphic in Ω.

First consider the set

$$U = \{ z \in \Omega; \quad e - T(z) \text{ is invertible } \} \neq \emptyset$$

U is open (by lemma 1 and continuity of $z \mapsto T(z)$) and $z \mapsto [e - T(z)]^{-1}$ is holomorphic in U (see lemma 6), hence meromorphic.

Now, suppose $z_0 \in \Omega \setminus U$. We are going to prove that, for all z in a neighborhood of z_0, the operator $e - T(z)$ can be splited into a regular part (with holomorphic inverse) and a singular part on a finite dimension space (hence with meromorphic inverse).

Since $e-t(z_0)$ is not invertible, we have $1 \in \sigma(T(z_0))$, where $\sigma(T(z_0))$ denote the spectrum of operator $T(z_0)$. Since $T(z_0)$ is compact, 1 is an isolated point of $\sigma(T(z_0))$ (see lemma 4). Thus we can choose O_1 and O_2, two open subsets of $\mathbb{C}$, such that

$$\left\{ \begin{array}{rcl} O_1 \cap O_2 & = & \emptyset \\ O_1 \cap \sigma(T(z_0)) & = & \{1\} \\ \sigma(T(z_0)) & \subset & O_1 \cup O_2 \end{array} \right. \tag{6}$$

Furthermore, we can choose two strictly smaller open subsets $O_1' \subset O_1$ and $O_2' \subset O_2$ with the same conditions (6). Now consider $\Gamma = \Gamma_1 \cup \Gamma_2$, a curve which surrounds $\sigma(T(z_0))$ in $(O_1 \cup O_2) \setminus (O_1' \cup O_2')$. See figure 1.

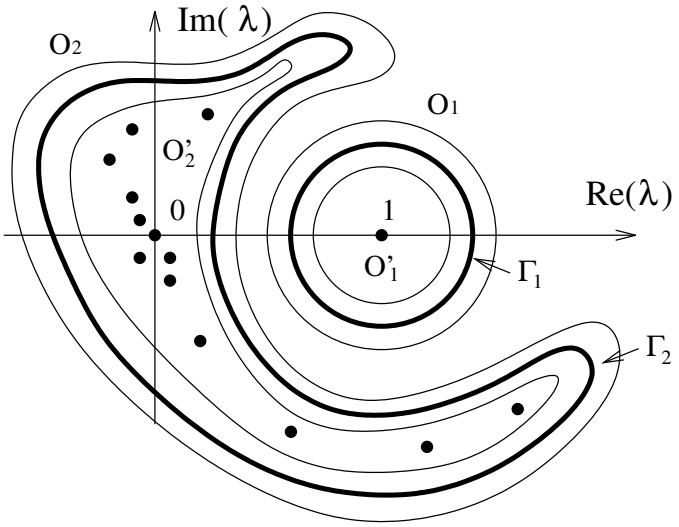


Figure 1: Spectrum of $T(z_0)$.

Due to lemma 2 and the continuity of $z \mapsto T(z)$ in z_0, there exists a connected (choose a polydisc, for example) neighborhood $V \subset \Omega$ of z_0 such that

$$\forall z \in V, \quad \sigma(T(z)) \subset O_1' \cup O_2' \tag{7}$$

and thus Γ surrounds $\sigma(T(z))$ in $(O_1 \cup O_2) \setminus (O_1' \cup O_2')$.
For all $z \in V$, we consider

$$P(z) = \int_{\Gamma_1} (\lambda e - T(z))^{-1} d\lambda = \widetilde{\mathbf{1}_{O_1}}(T(z)) \tag{8}$$

where $\mathbf{1}_{O_1}(\lambda) = 1$ if $\lambda \in O_1$ and $\mathbf{1}_{O_1}(\lambda) = 0$ if $\lambda \in O_2$ (note that $\mathbf{1}_{O_1}$ is therefore holomorphic in $O_1 \cup O_2$). By lemma 8, the mapping $z \mapsto P(z)$ is holomorphic in V.

Since $\mathbf{1}_{O_1}(\lambda)\mathbf{1}_{O_1}(\lambda) = \mathbf{1}_{O_1}(\lambda)$, the spectral mapping theorem (lemma 3) implies that

$$P(z)P(z) = P(z) \tag{9}$$

Let $M_z = \mathcal{R}(P)$ (range of P) and $M_z^c = \mathcal{N}(P)$ (null space of P). We have $M_z \cap M_z^c = \{0\}$ and $X = M_z + M_z^c$ (easy, see [5, p.126]). Moreover, we have

$$P(z)T(z) = T(z)P(z) \tag{10}$$

Indeed, since $T(z) = \widetilde{\lambda}(T(z))$ and $\lambda\mathbf{1}_{O_1}(\lambda) = \mathbf{1}_{O_1}(\lambda)\lambda$, the spectral mapping theorem gives again the result. Hence $T(z)M_z \subset M_z$ and $T(z)M_z^c \subset M_z^c$.

The key point is that $P(z)$ is compact. Indeed, for all $\lambda \in O_1$, we have $(\lambda e - T(z))^{-1}(\lambda e - T(z)) = e$. Since $0 \notin O_1$, $(\lambda e - T(z))^{-1} - \frac{1}{\lambda}e = \frac{1}{\lambda}(\lambda e - T(z))^{-1}T(z)$ is compact. Since $\lambda \mapsto \frac{1}{\lambda}e$ is holomorphic in O_1, we have $\int_{\Gamma_1} \frac{1}{\lambda}e d\lambda = 0$. Hence

$$P(z) = \int_{\Gamma_1} \left[(\lambda e - T(z))^{-1} - \frac{1}{\lambda}e \right] d\lambda \tag{11}$$

is compact as a limit of compact operators. Hence $M_z = \mathcal{N}(e - P(z))$ is a finite dimensional vector space (lemma 5).

However, it is no use to reduce directly $T(z)$ by the projection $P(z)$ because M_z depends on z. In this case it is usual to introduce

$$\begin{aligned} U(z) &= P(z_0)P(z) + (e - P(z_0))(e - P(z)) && (12) \\ &= e - [(P(z) - P(z_0))P(z) - P(z_0)(P(z) - P(z_0))] && (13) \end{aligned}$$

Since $z \mapsto P(z)$ is continuous, there exists a neighborhood V of z_0 such that

$$|(P(z) - P(z_0))P(z) - P(z_0)(P(z) - P(z_0))| < 1$$

for all $z \in V$. Thus $U(z)$ is invertible. Since it is obviously holomorphic in V (use lemma 7), so is $U(z)^{-1}$ (lemma 6). On the other hand, it is clear that, for all $z \in V$,

$$\begin{aligned} U(z)M_z &= M_{z_0} \\ U(z)M_z^c &= M_{z_0}^c \end{aligned} \tag{14}$$

Now consider the following operator (for all $z \in V$)

$$B(z) = U(z)\,[e - T(z)]\,U(z)^{-1} \tag{15}$$

$B(z)$ is splited into two parts.

$$B(z) = B_1(z)P(z_0) + B_2(z)(e - P(z_0)) \tag{16}$$

where :

- $B_1(z) = B(z)|_{M_{z_0}} : M_{z_0} \longrightarrow M_{z_0}$
 Since M_{z_0} is finite dimensional, we can use the Cramer's rule as in lemma 9. The only thing to prove is that $det(B_1(z))$ can not be identically null in V. Suppose the contrary. $B(z)$ is nowhere invertible in V. Thus $1 \in \sigma(T(z))$ for all $z \in V$. Consequently z_0 is an interior point of $\mathcal{P} = \{z \in \Omega :\ 1 \in \sigma(T(z))\}$. Thus $\dot{\mathcal{P}}$ is a non empty open subset of Ω. We are going to show that $\dot{\mathcal{P}}$ is also closed. Let z' be an accumulation point of $\dot{\mathcal{P}}$. There exist $z'_n \in \dot{\mathcal{P}}$ such that $z'_n \longrightarrow z'$ when $n \to +\infty$. Since $\mathcal{P}$ is closed, $z' \in \mathcal{P}$. The same construction as above can be done for z'. In particular, there exists a connected neighborhood V' of z' where $B'(t) = B'_1(t)P'(z') + B'_2(t)(e - P'(z'))$. There exists $n \in \mathbb{N}$ such that $z'_n \in V'$. Since $z'_n \in \dot{\mathcal{P}}$, there exists an open neighborhood W of z'_n such that $W \subset \mathcal{P}$. Since $B'_2(t)$ is invertible (see below), $B'_1(t)$ is not. Hence $det(B'_1(t)) \equiv 0$ in $W \cap V'$, thus in V' because V' is connected. This proves $z' \in \dot{\mathcal{P}}$. Since $\dot{\mathcal{P}}$ is a non empty open and closed subset of the connected set Ω, we have $\dot{\mathcal{P}} = \Omega$, which is in contradiction with the hypothesis $(e - T(z))$ invertible in $\mathcal{B}(X)$ for some $z \in \Omega$.
 Thus $det(B_1(z)) \not\equiv 0$ in V and $z \mapsto B_1(z)^{-1}$ is meromorphic in V.

- $B_2(z)|_{M^c_{z_0}} : M^c_{z_0} \longrightarrow M^c_{z_0}$
 We have $B_2(z)x = U(z)\left[e - T(z)(e - P(z))\right]U(z)^{-1}x$, for all $x \in M^c_{z_0}$. Since we have $T(z)(e - P(z)) = \widetilde{\lambda \mathbf{1}_{O_2}}(T(z))$, the spectral mapping theorem gives $\sigma(T(z)(e - P(z))) = (\lambda \mathbf{1}_{O_2})(\sigma(T(z))$, hence $1 \notin \sigma(T(z)(e - P(z)))$. Thus $[e - T(z)(e - P(z))]$ is invertible, and so is $B_2(z)$. Thus $z \mapsto B_2(z)^{-1}$ is holomorphic in V (lemma 6).

Hence $z \mapsto B(z)^{-1} = B_1(z)^{-1}P(z_0) + B_2(z)^{-1}(e - P(z_0))$ is meromorphic in V. Thus so is $z \mapsto [e - T(z)]^{-1} = U(z)B(z)^{-1}U(z)^{-1}$ and this ends the proof of theorem 1.

6 Final comments

In the above proof we have shown that the interior of the polar set $\mathcal{P} = \{z \in \Omega :\ 1 \in \sigma(T(z))\}$ is empty. On the other hand, $\mathcal{P}$ is an analytic set and has other properties not discussed here (for example, see [3]).

In the end let us mention some links between Steinberg alternative and Fredholm's one.

Fredholm alternative. *Let $T \in \mathcal{B}(X)$ be compact. Then the following alternative holds :*

(i) either $e - T$ is invertible (hence $\mathcal{N}(e - T) = \{0\}$ and equation $x - Tx = f$ has a unique solution for all f);

(ii) either $0 < dim\mathcal{N}(e - T) < \infty$ (equation $x - Tx = f$ has infinitly many solutions which form a vector space of dimension $dim\mathcal{N}(e - T)$ if $f \in \mathcal{R}(e - T)$, and has no solution if not.).

Both are easy in finite dimension, and they use compacity for running in infinite dimension. Now consider a holomorphic compact operator-valued function $z \mapsto T(z)$ in a connected set $\Omega \subset \mathbb{C}^d$. If $e - T(z)$ is nowhere invertible in Ω (case (i) of Steinberg alternative) then case (ii) of Fredholm alternative holds. Or else $(e - T(z))^{-1}$ is meromorphic in Ω (case (ii) of Steinberg alternative). Hence, for every $z_0 \in \Omega$, either $z \mapsto (e - T(z))^{-1}$ is holomorphic in a neighborhood of z_0, where case (i) of Fredholm alternative holds, either $e - T(z_0)$ is not invertible and case (ii) of Fredholm alternative holds.

References

[1] D. Boichu, V. Robin, Multivariate Padé-Bergman Approximants, *Numerische Mathematik*, **93**, p. 223-238, 2002.

[2] A. Huard, V. Robin, Continuity of Approximation by Least Squares Multivariate Padé Approximants, *Journal of Computational and Applied Mathematics,* **75**, p. 255-268, 2000.

[3] R.M. Range, *Holomorphic functions and integral representation in several complex variables*, Springer-Verlag, New-York, 1986.

[4] V. Robin, *Utilisation des dérivées d'ordre élevé pour le calcul en électromagnétisme*, Thèse de doctorat, Université Paul Sabatier, Toulouse, France, 1997.

[5] W. Rudin, *Functional Analysis*, TMH Edition, 1974.

[6] S. Steinberg, Meromorphic Families of Compact Operators, *Arch. Rational Mech. Anal.*, 31 (1968), pp. 372–379.

Instructions to Contributors
Journal of Concrete and Applicable Mathematics

A quartely international publication of Eudoxus Press, LLC, of TN.

Editor in Chief: George Anastassiou
Department of Mathematical Sciences
University of Memphis
Memphis, TN 38152-3240, U.S.A.

1. Manuscripts hard copies in triplicate, and in English, should be submitted to the Editor-in-Chief:

Prof.George A. Anastassiou
Department of Mathematical Sciences
The University of Memphis
Memphis,TN 38152, USA.
Tel.001. 901.678.3144
e-mail:ganastss@memphis.edu.

Authors may want to recommend an associate editor the most related to the submission to possibly handle it.
Also authors may want to submit a list of six possible referees, to be used in case we cannot find related referees by ourselves.
2. Manuscripts should be typed using any of TEX,LaTEX,AMS-TEX,or AMS-LaTEX and according to EUDOXUS PRESS, LLC. LATEX STYLE FILE. (VISIT www.msci.memphis.edu/~ganastss/jcaam / to save a copy of the style file.)They should be carefully prepared in all respects. Submitted copies should be brightly printed (not dot-matrix), double spaced, in ten point type size, on one side high quality paper 8(1/2)x11 inch. Manuscripts should have generous margins on all sides and should not exceed 24 pages.
3. Submission is a representation that the manuscript has not been published previously in this or any other similar form and is not currently under consideration for publication elsewhere. A statement transferring from the authors(or their employers,if they hold the copyright) to Eudoxus Press, LLC, will be required before the manuscript can be accepted for publication.The Editor-in-Chief will supply the necessary forms for this transfer.Such a written transfer of copyright,which previously was assumed to be implicit in the act of submitting a manuscript,is necessary under the U.S.Copyright Law in order for the publisher to carry through the dissemination of research results and reviews as widely and effective as possible.
4. The paper starts with the title of the article, author's name(s) (no titles or degrees), author's affiliation(s) and e-mail addresses. The affiliation should comprise the department, institution (usually university or company), city, state (and/or nation) and mail code.
The following items, 5 and 6, should be on page no. 1 of the paper.

5. An abstract is to be provided, preferably no longer than 150 words.
6. A list of 5 key words is to be provided directly below the abstract. Key words should express the precise content of the manuscript, as they are used for indexing purposes. The main body of the paper should begin on page no. 1, if possible.
7. All sections should be numbered with Arabic numerals (such as: 1. INTRODUCTION) .
Subsections should be identified with section and subsection numbers (such as 6.1. Second-Value Subheading).
If applicable, an independent single-number system (one for each category) should be used to label all theorems, lemmas, propositions, corrolaries, definitions, remarks, examples, etc. The label (such as Lemma 7) should be typed with paragraph indentation, followed by a period and the lemma itself.
8. Mathematical notation must be typeset. Equations should be numbered consecutively with Arabic numerals in parentheses placed flush right,and should be thusly referred to in the text [such as Eqs.(2) and (5)]. The running title must be placed at the top of even numbered pages and the first author's name, et al., must be placed at the top of the odd numbed pages.
9. Illustrations (photographs, drawings, diagrams, and charts) are to be numbered in one consecutive series of Arabic numerals. The captions for illustrations should be typed double space. All illustrations, charts, tables, etc., must be embedded in the body of the manuscript in proper, final, print position. In particular, manuscript, source, and PDF file version must be at camera ready stage for publication or they cannot be considered.
Tables are to be numbered (with Roman numerals) and referred to by number in the text. Center the title above the table, and type explanatory footnotes (indicated by superscript lowercase letters) below the table.
10. List references alphabetically at the end of the paper and number them consecutively. Each must be cited in the text by the appropriate Arabic numeral in square brackets on the baseline.

References should include (in the following order):
initials of first and middle name, last name of author(s)
title of article,
name of publication, volume number, inclusive pages, and year of publication.
Authors should follow these examples:

Journal Article

1. H.H.Gonska,Degree of simultaneous approximation of bivariate functions by Gordon operators, (journal name in italics) *J. Approx. Theory*, 62,170-191(1990).

Book

2. G.G.Lorentz, (title of book in italics) *Bernstein Polynomials* (2nd ed.), Chelsea,New York,1986.

Contribution to a Book

3. M.K.Khan, Approximation properties of beta operators,in(title of book in italics) *Progress in Approximation Theory* (P.Nevai and A.Pinkus,eds.), Academic Press, New York,1991,pp.483-495.

11. All acknowledgements (including those for a grant and financial support) should occur in one paragraph that directly precedes the References section.
12. Footnotes should be avoided. When their use is absolutely necessary, footnotes should be numbered consecutively using Arabic numerals and should be typed at the bottom of the page to which they refer. Place a line above the footnote, so that it is set

off from the text. Use the appropriate superscript numeral for citation in the text.
13. After each revision is made please again submit three hard copies of the revised manuscript, including in the final one. And after a manuscript has been accepted for publication and with all revisions incorporated, manuscripts, including the TEX/LaTex source file and the PDF file, are to be submitted to the Editor's Office on a personal-computer disk, 3.5 inch size. Label the disk with clearly written identifying information and properly ship, such as:

Your name, title of article, kind of computer used, kind of software and version number, disk format and files names of article, as well as abbreviated journal name.

Package the disk in a disk mailer or protective cardboard. Make sure contents of disks are identical with the ones of final hard copies submitted!

Note: The Editor's Office cannot accept the disk without the accompanying matching hard copies of manuscript. No e-mail final submissions are allowed! The disk submission must be used.

14. Effective 1 Nov. 2005 the journal's page charges are $8.00 per PDF file page. Upon acceptance of the paper an invoice will be sent to the contact author. The fee payment will be due one month from the invoice date. The article will proceed to publication only after the fee is paid. The charges are to be sent, by money order or certified check, in US dollars, payable to Eudoxus Press, LLC, to the address shown in the Scope and Prices section.

No galleys will be sent and the contact author will receive one(1) electronic copy of the journal issue in which the article appears.

15. This journal will consider for publication only papers that contain proofs for their listed results.

TABLE OF CONTENTS,JOURNAL OF CONCRETE AND APPLICABLE MATHEMATICS,VOL.3,NO.1,2005

HERMITE-KAMPE DE FERIET POLYNOMIALS AND SOLUTIONS OF BOUNDARY VALUE PROBLEMS IN THE HALF-SPACE,
G.MAROSCIA,P.RICCI,…….9

GENERALIZED m-ACCRETIVE MAPPINGS AND VARIATIONAL INCLUSIONS IN BANACH SPACES,
N.HUANG,Y.FANG,Y.CHO,…….31

C1,1 FUNCTIONS AND OPTIMALITY CONDITIONS,
D.LaTORRE,M.ROCCA,……....41

NUMERICAL ANALYSIS OF A QUASISTATIC SLIDING CONTACT PROBLEM WITH WEAR,J.FERNANDEZ,M.SOFONEA,J.VIANO,……..55

CHARACTERIZATION OF A CLASS OF BAND-LIMITED WAVELETS,
B.BEHERA,S.MADAN,……..75

OSTROWSKI TYPE INEQUALITIES FOR FUNCTIONS DEFINED ON LINEAR SPACES AND APPLICATIONS FOR SEMI-INNER PRODUCTS,
S.DRAGOMIR,…….91

MULTIVARIATE STEINBERG THEOREM,D.BOICHU,V.ROBIN,……..105

VOLUME 3,NUMBER 2 APRIL 2005

ISSN:1548-5390 PRINT,1559-176X ONLINE



JOURNAL

OF CONCRETE

AND APPLICABLE

MATHEMATICS

EUDOXUS PRESS,LLC

SCOPE AND PRICES OF THE JOURNAL
Journal of Concrete and Applicable Mathematics

A quartely international publication of **Eudoxus Press,LLC**

Editor in Chief: George Anastassiou
Department of Mathematical Sciences,
University of Memphis
Memphis, TN 38152, U.S.A.
ganastss@memphis.edu

The main purpose of the "Journal of Concrete and Applicable Mathematics" is to publish high quality original research articles from all subareas of Non-Pure and/or Applicable Mathematics and its many real life applications, as well connections to other areas of Mathematical Sciences, as long as they are presented in a Concrete way. It welcomes also related research survey articles and book reviews.A sample list of connected mathematical areas with this publication includes and is not restricted to: Applied Analysis, Applied Functional Analysis, Probability theory, Stochastic Processes, Approximation Theory, O.D.E, P.D.E, Wavelet, Neural Networks,Difference Equations, Summability, Fractals, Special Functions, Splines, Asymptotic Analysis, Fractional Analysis, Inequalities, Moment Theory, Numerical Functional Analysis,Tomography, Asymptotic Expansions, Fourier Analysis, Applied Harmonic Analysis, Integral Equations, Signal Analysis, Numerical Analysis, Optimization, Operations Research, Linear Programming, Fuzzyness, Mathematical Finance, Stochastic Analysis, Game Theory, Math.Physics aspects, Applied Real and Complex Analysis, Computational Number Theory, Graph Theory, Combinatorics,
Computer Science Math.related topics,combinations of the above, etc. In general any kind of Concretely presented Mathematics which is Applicable fits to the scope of this journal.
Working Concretely and in Applicable Mathematics has become a main trend in many recent years,so we can understand better and deeper and solve the important problems of our real and scientific world.
"Journal of Concrete and Applicable Mathematics" is a peer- reviewed International Quarterly Journal.
We are calling for papers for possible publication. The contributor should send three copies of the contribution to the editor in-Chief typed in TEX, LATEX double spaced. [See: Instructions to Contributors]

Journal of Concrete and Applicable Mathematics(JCAAM)
ISSN:1548-5390 PRINT, 1559-176X ONLINE.
is published in January,April,July and October of each year by
EUDOXUS PRESS,LLC,
1424 Beaver Trail Drive,Cordova,TN38016,USA,
Tel.001-901-751-3553
anastassioug@yahoo.com
http://www.EudoxusPress.com.

Annual Subscription Current Prices:For USA and Canada,Institutional:Print $250,Electronic $220,Print and Electronic $310.Individual:Print $77,Electronic $60,Print &Electronic $110.For any other part of the world add $25 more to the above prices for Print.

Single article PDF file for individual $8.Single issue in PDF form for individual $25.
The journal carries page charges $8 per page of the pdf file of an article,payable upon acceptance of the article within one month and before publication.
No credit card payments.Only certified check,money order or international check in US dollars are acceptable.
Combination orders of any two from JoCAAA,JCAAM,JAFA receive 25% discount,all three receive 30% discount.

Copyright©2004 by Eudoxus Press,LLC all rights reserved.JCAAM is printed in USA.

JCAAM is reviewed and abstracted by AMS Mathematical Reviews,MATHSCI,and Zentralblaat MATH.

It is strictly prohibited the reproduction and transmission of any part of JCAAM and in any form and by any means without the written permission of the publisher.It is only allowed to educators to Xerox articles for educational purposes.The publisher assumes no responsibility for the content of published papers.
JCAAM IS A JOURNAL OF RAPID PUBLICATION

Editorial Board
Associate Editors

Editor in -Chief:
George Anastassiou
Department of Mathematical Sciences
The University Of Memphis
Memphis,TN 38152,USA
tel.901-678-3144,fax 901-678-2480
e-mail ganastss@memphis.edu
www.msci.memphis.edu/~ganastss/jcaam
Areas:Approximation Theory,
Probability,Moments,Wavelet,
Neural Networks,Inequalities,Fuzzyness.

Associate Editors:

1) Ravi Agarwal
Florida Institute of Technology
Applied Mathematics Program
150 W.University Blvd.
Melbourne,FL 32901,USA
agarwal@fit.edu
Differential Equations,Difference
Equations,inequalities

2) Shair Ahmad
University of Texas at San Antonio
Division of Math.& Stat.
San Antonio,TX 78249-0664,USA
SAHMAD@utsa.edu
Differential Equations,Mathematical
Biology

3) Drumi D.Bainov
Medical University of Sofia
P.O.Box 45,1504 Sofia,Bulgaria
drumibainov@yahoo.com
Differential Equations,Optimal Control,
Numerical Analysis,Approximation Theory

4) Carlo Bardaro
Dipartimento di Matematica & Informatica
Universita' di Perugia
Via Vanvitelli 1
06123 Perugia,ITALY
tel.+390755855034, +390755853822,
fax +390755855024
bardaro@unipg.it ,
bardaro@dipmat.unipg.it
Functional Analysis and Approximation Th.,

19) Rupert Lasser
Institut fur Biomathematik & Biomertie,GSF
-National Research Center for environment and
health
Ingolstaedter landstr.1
D-85764 Neuherberg,Germany
lasser@gsf.de
Orthogonal Polynomials,Fourier Analysis,
Mathematical Biology

20) Alexandru Lupas
University of Sibiu
Faculty of Sciences
Department of Mathematics
Str.I.Ratiu nr.7
2400-Sibiu,Romania
lupas@ulbsibiu.ro
Classical Analysis,Inequalities,
Special Functions,Umbral Calculus,
Approximation Th.,Numerical Analysis
and Methods

21) Ram N.Mohapatra
Department of Mathematics
University of Central Florida
Orlando,FL 32816-1364
tel.407-823-5080
ramm@mail.ucf.edu
Real and Complex analysis,Approximation Th.,
Fourier Analysis, Fuzzy Sets and Systems

22) Rainer Nagel
Arbeitsbereich Funktionalanalysis
Mathematisches Institut
Auf der Morgenstelle 10
D-72076 Tuebingen
Germany
tel.49-7071-2973242
fax 49-7071-294322
rana@fa.uni-tuebingen.de
evolution equations,semigroups,spectral th.,
positivity

23) Panos M.Pardalos
Center for Appl. Optimization
University of Florida
303 Weil Hall
P.O.Box 116595
Gainesville,FL 32611-6595

Summability,Signal Analysis,Integral Equations,
Measure Th.,Real Analysis

5) Francoise Bastin
Institute of Mathematics
University of Liege
4000 Liege
BELGIUM
f.bastin@ulg.ac.be
Functional Analysis,Wavelets

6) Paul L.Butzer
RWTH Aachen
Lehrstuhl A fur Mathematik
D-52056 Aachen
Germany
tel.0049/241/80-94627 office,
0049/241/72833 home,
fax 0049/241/80-92212
Butzer@rwth-aachen.de
Approximation Th.,Sampling Th.,Signals,
Semigroups of Operators,Fourier Analysis

7) Yeol Je Cho
Department of Mathematics Education
College of Education
Gyeongsang National University
Chinju 660-701
KOREA
tel.055-751-5673 Office,
055-755-3644 home,
fax 055-751-6117
yjcho@nongae.gsnu.ac.kr
Nonlinear operator Th.,Inequalities,
Geometry of Banach Spaces

8) Sever S.Dragomir
School of Communications and Informatics
Victoria University of Technology
PO Box 14428
Melbourne City M.C
Victoria 8001,Australia
tel 61 3 9688 4437,fax 61 3 9688 4050
sever.dragomir@vu.edu.au,
sever@sci.vu.edu.au
Math.Analysis,Inequalities,Approximation Th.,
Numerical Analysis, Geometry of Banach Spaces,
Information Th. and Coding

9) A.M.Fink
Department of Mathematics
Iowa State University
Ames,IA 50011-0001,USA
tel.352-392-9011
pardalos@ufl.edu
Optimization,Operations Research

24) Svetlozar T.Rachev
Dept.of Statistics and Applied Probability Program
University of California,Santa Barbara
CA 93106-3110,USA
tel.805-893-4869
rachev@pstat.ucsb.edu
AND
Chair of Econometrics and Statistics
School of Economics and Business Engineering
University of Karlsruhe
Kollegium am Schloss,Bau II,20.12,R210
Postfach 6980,D-76128,Karlsruhe,Germany
tel.011-49-721-608-7535
rachev@lsoe.uni-karlsruhe.de
Mathematical and Empirical Finance,
Applied Probability, Statistics and Econometrics

25) Paolo Emilio Ricci
Universita' degli Studi di Roma "La Sapienza"
Dipartimento di Matematica-Istituto "G.Castelnuovo"
P.le A.Moro,2-00185 Roma,ITALY
tel.++39 0649913201,fax ++39 0644701007
riccip@uniroma1.it,Paoloemilio.Ricci@uniroma1.it
Orthogonal Polynomials and Special functions,
Numerical Analysis, Transforms,Operational Calculus,
Differential and Difference equations

26) Cecil C.Rousseau
Department of Mathematical Sciences
The University of Memphis
Memphis,TN 38152,USA
tel.901-678-2490,fax 901-678-2480
ccrousse@memphis.edu
Combinatorics,Graph Th.,
Asymptotic Approximations,
Applications to Physics

27) Tomasz Rychlik
Institute of Mathematics
Polish Academy of Sciences
Chopina 12,87100 Torun, Poland
T.Rychlik@impan.gov.pl
Mathematical Statistics,Probabilistic Inequalities

28) Bl. Sendov
Institute of Mathematics and Informatics
Bulgarian Academy of Sciences
Sofia 1090,Bulgaria

tel.515-294-8150
fink@math.iastate.edu
Inequalities,Ordinary Differential Equations

10) Sorin Gal
Department of Mathematics
University of Oradea
Str.Armatei Romane 5
3700 Oradea,Romania
galso@uoradea.ro
Approximation Th.,Fuzzyness,Complex Analysis

11) Jerome A.Goldstein
Department of Mathematical Sciences
The University of Memphis,
Memphis,TN 38152,USA
tel.901-678-2484
jgoldste@memphis.edu
Partial Differential Equations,
Semigroups of Operators

12) Heiner H.Gonska
Department of Mathematics
University of Duisburg
Duisburg,D-47048
Germany
tel.0049-203-379-3542 office
gonska@informatik.uni-duisburg.de
Approximation Th.,Computer Aided Geometric Design

13) Dmitry Khavinson
Department of Mathematical Sciences
University of Arkansas
Fayetteville,AR 72701,USA
tel.(479)575-6331,fax(479)575-8630
dmitry@uark.edu
Potential Th.,Complex Analysis,Holomorphic PDE,Approximation Th.,Function Th.

14) Virginia S.Kiryakova
Institute of Mathematics and Informatics
Bulgarian Academy of Sciences
Sofia 1090,Bulgaria
virginia@diogenes.bg
Special Functions,Integral Transforms,
Fractional Calculus

15) Hans-Bernd Knoop
Institute of Mathematics
Gerhard Mercator University
D-47048 Duisburg
Germany
tel.0049-203-379-2676
bsendov@bas.bg
Approximation Th.,Geometry of Polynomials,
Image Compression

29) Igor Shevchuk
Faculty of Mathematics and Mechanics
National Taras Shevchenko
University of Kyiv
252017 Kyiv
UKRAINE
shevchuk@univ.kiev.ua
Approximation Theory

30) H.M.Srivastava
Department of Mathematics and Statistics
University of Victoria
Victoria,British Columbia V8W 3P4
Canada
tel.250-721-7455 office,250-477-6960 home,
fax 250-721-8962
harimsri@math.uvic.ca
Real and Complex Analysis,Fractional Calculus and Appl.,
Integral Equations and Transforms,Higher Transcendental
Functions and Appl.,q-Series and q-Polynomials,
Analytic Number Th.

31) Ferenc Szidarovszky
Dept.Systems and Industrial Engineering
The University of Arizona
Engineering Building,111
PO.Box 210020
Tucson,AZ 85721-0020,USA
szidar@sie.arizona.edu
Numerical Methods,Game Th.,Dynamic Systems,
Multicriteria Decision making,
Conflict Resolution,Applications
in Economics and Natural Resources Management

32) Gancho Tachev
Dept.of Mathematics
Univ.of Architecture,Civil Eng. and Geodesy
1 Hr.Smirnenski blvd
BG-1421 Sofia,Bulgaria
Approximation Theory

33) Manfred Tasche
Department of Mathematics
University of Rostock
D-18051 Rostock
Germany
manfred.tasche@mathematik.uni-rostock.de
Approximation Th.,Wavelet,Fourier Analysis,
Numerical Methods,Signal Processing,

knoop@math.uni-duisburg.de
Approximation Theory,Interpolation

16) Jerry Koliha
Dept. of Mathematics & Statistics
University of Melbourne
VIC 3010,Melbourne
Australia
koliha@unimelb.edu.au
Inequalities,Operator Theory,
Matrix Analysis,Generalized Inverses

17) Mustafa Kulenovic
Department of Mathematics
University of Rhode Island
Kingston,RI 02881,USA
kulenm@math.uri.edu
Differential and Difference Equations

18) Gerassimos Ladas
Department of Mathematics
University of Rhode Island
Kingston,RI 02881,USA
gladas@math.uri.edu
Differential and Difference Equations
Image Processing,Harmonic Analysis

34) Chris P.Tsokos
Department of Mathematics
University of South Florida
4202 E.Fowler Ave.,PHY 114
Tampa,FL 33620-5700,USA
profcpt@math.usf.edu,profcpt@chuma1.cas.usf.edu
Stochastic Systems,Biomathematics,
Environmental Systems,Reliability Th.

35) Lutz Volkmann
Lehrstuhl II fuer Mathematik
RWTH-Aachen
Templergraben 55
D-52062 Aachen
Germany
volkm@math2.rwth-aachen.de
Complex Analysis,Combinatorics,Graph Theory

JOURNAL OF CONCRETE AND APPLICABLE MATHEMATICS,VOL.3,NO.2,127-168,2005,COPYRIGHT 2005 EUDOXUS PRESS,LLC

ON A NEW FAMILY OF GAUSSIAN QUADRATURE FORMULAE OF BIRKHOFF TYPE WITH APPLICATIONS TO POLYNOMIAL INEQUALITIES

ALLAL GUESSAB

Abstract. A new family of Gauss type quadrature formulae, which contain terms involving certain derivatives at an interior point in the interval of integration, is first presented and discussed. It is then shown how these quadratures can be constructed via eigenvalues and eigenvectors of a real symmetric tridiagonal matrix. Numerical comparisons are made with the classical Gaussian quadrature formulae. We also derive new explicit sharp weighted L^2 inequalities for coefficients of polynomials. They are all obtained as direct applications of the quadrature formulae created in this paper.

1. Introduction and motivations

The aim of this series of papers is to propose and analyze methods for developing efficient algorithms for the numerical computation of a new class of quadrature formulae. The discussion of [17] and [27] is partly concerned with quadratures that contain boundary terms involving some derivatives at both endpoints of the interval of integration. These quadrature formulae have the novel feature that they can best be appreciated in the case of integrals with severely ill-behaved integrands such as those with singularities at one or both *endpoint*. However, such techniques or popular methods are often computationally inefficient when integrating a function that has both "interior" regions with large functional variation and regions with small functional variation. This is entirely due to the excessive number of points required to achieve a reasonable amount of accuracy. In this part of the series, we shall present and construct a different class of quadratures which uses terms involving some well-determined derivatives of the integrand at certain sampling *interior* nodes fixed in advance and suitably chosen in the interval of integration. The remaining (free) nodes and weights are determined so as enable a minimum number of function evaluations.

The choice of an appropriate quadrature for solving a numerical integration problem depends on the form and content of helpful singularity information about the integrand -like important information about the values of specific points or bounds of certain derivatives-. The most effective way we found for exploiting such additional informations is to incorporate in the quadrature formula some of these values, that can be chosen judiciously at locations where the integrand function is predominant. Thus, this approach may indeed have a strong influence on the

1991 *Mathematics Subject Classification.* Primary 42C05, 65D07, 65D15, 65D30, 65D32.

Key words and phrases. Quadrature formulae, Gaussian quadrature formulae, Jacobi polynomials, quasi-orthogonal polynomials, three-term relation, algorithms.

This research was supported by the CNRS under grant INTAS-94-4070.

accurately and efficiently of such quadratures, because, as already expressed by Christoffel (cf. [18, p. 86]), it often helps to find quadratures extremely well suited for numerical treatment in practice and it gives good ways of taking effective use of available information about certain smoothness properties of the integrand.

The main purpose of this paper is centered around the development of a simple procedure for computing numerically the nodes $x_{l,m,n} = x_{l,m,n}(d\sigma; j_1, ..., j_m)$ and the weights $\lambda_{l,m,n} = \lambda_{l,m,n}(d\sigma; j_1, ..., j_m)$ and $\omega_{j_i,m,n} = \omega_{j_i,m,n}(d\sigma; j_1, ..., j_m)$, such that the new quadrature formula,

$$Q_{\sigma,m,n}(f; j_1, ..., j_m) = \sum_{i=1}^{m} \omega_{j_i,m,n} f^{(2j_i)}(0) + \sum_{l=1}^{n} \lambda_{l,m,n} f(x_{l,m,n}), \tag{1.1}$$

has maximum polynomial degree of exactness (MPDE)

$$MPDE(Q_{\sigma,m,n}) = 2n + 2m - 1,$$

i. e.,

$$\begin{aligned} I_\sigma(f) &= \int_{-1}^{1} f(x)\, d\sigma(x) \\ &= Q_{\sigma,m,n}(f; j_1, ..., j_m), \quad \forall f \in \mathcal{P}_{2n+2m-1}. \end{aligned}$$

Here and subsequently $j_1, ..., j_m$, are given distinct real arranged in increasing order such that $2j_i \leq 2n + 2m - 1$, $i = 1, ..., m$. For simplicity of presentation, we shall suppose that $d\sigma$ is a even (nonnegative) measure with symmetric support with respect to the origin on the interval $[-1, 1]$ and all of whose moments exist. It is easy to modify the results to cover the more general case of an arbitrary nonnegative measure. Note that, these formulae share the property of the classical Gaussian formula that a formula with $N = n + m$ evaluations is exact for all polynomials of degree $\leq 2N - 1$. Thus, they can be viewed as natural extension of the standard Gaussian quadrature.

The problem stated, in the special case $j_i = i$, $i = 1, ..., m$, is classical, and the best of our knowledge there are only two methods for computing a such type of quadrature. One of them [41] is based on solving by Newton's method the nonlinear moment equations expressing exactness of the formula. Such a procedure is subject to considerable loss of accuracy and therefore required elevated precision. This is considered its main disadvantage. The other [24] is a very general algorithm that uses among other tools the full eigenvector matrix of a tridiagonal matrix derived by matrix decomposition methods. The weights are then obtained using general methods for constructing interpolatory quadrature formulae [15]. Except for their availability only for small n, both these methods lack the rapid convergence and elegance of the well-known algorithms of Golub and Welch [23] and Gautschi [20, p. 86] for ordinary Gaussian case (case where preassigned nodes do not exist, i. e., $m = 0$). Apart from these rather *Hermitian* cases, surprisingly no effective methods is currently available to compute *Birkhoff-Hermite* type case.

For these reasons, we approach the problem of constructing the functional $Q_{k,n}$ differently. Here, our principal results give a numerical algorithm which offers an efficient way of reducing the computation of the functional $Q_{\sigma,m,n}$ (i. e., the computation of the nodes $x_{l,m,n}$ and weights $\lambda_{l,m,n}$) to the well-studied problem of computing ordinary Gaussian quadrature formulae from recurrence coefficients, and can therefore be brought into the realm of stable modern methods of constructing orthogonal polynomials; see Gautschi [18]. The only overhead of this computation is the construction of new Jacobi matrix by changing some recurrence coefficients

which characterize the original measure $d\sigma$. This is a crucial aspect of the method, since both the programming task and computational effort are elevated considerably.

The approach on this topic was stimulated by a request from a theoretical physicist to find a good quadrature which contains some derivatives at given interior nodes. However, the approach and techniques used here, which extend some of the results in [7], have been strongly motivated by the problem of Turán [43, Problem XXXIII], who asked if there exists a quadrature formula of the form

$$\int_{-1}^{1} f(x)d\sigma(x) \approx \sum_{i=1}^{n}(\lambda_i f(x_i) + \delta_i f^{(2k)}(x_i)), \tag{1.2}$$

which is exact for all $f \in \mathcal{P}_{2n}$. In this form the problem has been solved, in the case $k = 1$, by Dimitrov [12] and extended, for arbitrary k, by the same author in [13] constructing a quadrature formula which uses exactly $2n$ function evaluations. It is naturally led to ask whether a quadrature formula of type (1.2) might have a minimal number of function evaluations. It should be observed that since $d\sigma$ is even and the support of $d\sigma$ is symmetric with respect to the origin then the free nodes of (1.1) are located symmetrically in the interval $(-1, 1)$. Consequently, when n is odd then, due the symmetry, one of the nodes will coincide with the origin. Thus, if $m = 1$, the present work offers a simple solution for Turán problem. We also remark that not only the latter is more pleasing than those presented in [12, 13] since it requires considerably few evaluations ($n + 1$ instead of $2n$) and it can be computed most conveniently as eigenvalues of a symmetric tridiagonal matrix, but additionally it achieves the highest possible polynomial exactness among all nested quadrature formulae which use the same number of function evaluations.

Furthermore, this new class of quadrature formulae can find a wider range of applicability than merely to the particular type of problems considered here. In fact, in many physical applications it is not uncommon that no only the values of a function are accessible, but also some of its derivatives. In such a situation, it is meaningful to employ quadrature formulae involving these derivative values. An important method in which such type of quadratures can be applied in natural way is in the numerical solution of partial differential equations by spectral approximation. See, in this connection, Bernadi and Maday [4] where generalized Lobatto-type formulae have been successfully used, and the author's joint work [17] where efficient spectral methods based on a new family of quadrature formulae where complex boundary conditions terms were included.

The remaining six sections of this paper are organized as follows: In the next section, we prove existence and uniqueness of the quadratures we are seeking (1.1) and we also develop some of their properties. Then, in section 3, we give a detailed description of their characterization, and after that we present some properties of the Jacobi matrix associated with (1.1). The proofs are constructive, so it is straightforward to develop an algorithm which can compute (1.1). In section 4, we apply these properties together with the characterization results of section 3 to construct a class of quadrature formula of type (1.1), we also solve a classical integration problem of Turán [43, Problem XXXIII]. Numerical examples demonstrating the efficiency and accuracy of our quadrature formulae, as well as some comparisons with the classical Gauss quadrature formula are presented in section 5. Section 6 offers some sharp L^2 inequalities for coefficients of polynomials. These

inequalities are all obtained as direct applications of the quadrature formulae created in the previous sections. Finally, we give some concluding remarks, followed by an appendix in which we include some of more technically involved proofs of Section 5 and 6.

Remark 1. *Throughout this paper, we will take the "standardized" integration interval* $[-1,1]$ *as a matter of convenience, since any finite interval* $[a,b]$ *can be mapped onto* $[-1,1]$ *via an affine transformation and the resulting transformed integration formula has the same MPDE. To simplify the presentation, we only seek quadrature formulae which contain terms involving some derivatives at* 0*. The reader should bear in mind that it is, in particular, for integrands, with a (integrable) singularity or with a high peak at* 0 *or near* 0 *as well as functions that are highly oscillatory in a neighborhood of* 0*, that this quadrature formula is designed specifically. When the integrands have several peaks or an arbitrary number of internal point singularity, we suggest to subdivide the interval* $[-1,1]$ *into several subintervals, in such a way that the integrand has a peak or a singularity at the midpoint of each subinterval, and applying the quadrature formula (1.1) (using, if necessary, an appropriate linear transformation) to each subinterval separately. A consequence of this approach is that the most significant part of the analysis need only be carried out on the reference domain.*

2. Existence of the quadratures and some of their properties

Our first result states that there exists one and only one quadrature formula of type (1.1). It is also shown that (1.1) possesses most of the desirable properties of the classical polynomial Gaussian quadrature formulae, namely, the interlacing property of the free nodes, their inclusion in the support of the measure, and even more importantly the positivity of all weights $\lambda_{l,m,n}$. These, as is known classically, are important properties that numerical quadrature formulae required to have.

We first consider the problem of the existence and uniqueness of (1.1).

Theorem 1. *Given a nonnegative even measure* $d\sigma$ *and nonnegative integers* n, m, $j_1, j_2, ..., j_m$ *with* $n \geq 1$*, then there is precisely one quadrature formula of type (1.1), which integrates exactly all polynomial of* $\mathcal{P}_{2n+2m-1}$*. The nodes* $x_{1,m,n}, \ldots, x_{n,m,n}$ *are all in the open interval* $(-1,1)$*, and their weights* $\lambda_{1,m,n}, \ldots, \lambda_{n,m,n}$ *are positive.*

Proof. The proof is adapted version of the proof of Theorem 1 of the author's joint work [7].

I) *Uniqueness.* Suppose $z = \{z_{1,m,n}, ..., z_{n,m,n}\}$ and $y = \{y_{1,m,n}, ..., y_{n,m,n}\}$ were both sets of nodes such that the quadrature formula (1.1) based on each of them were exact for all polynomials from $\mathcal{P}_{2n+2m-1}$. Suppose that there exists $i_0 \in \{1, ..., n\}$ such that $z_{i_0,m,n} \neq y_{i_0,m,n}$. We now show that this impossible. In fact, by the well-known Atkinson-Sharma interpolation theorem [2], there exists a polynomial $P_{i_0} \in \mathcal{P}_{2n+2m-1}$ such that

$$\begin{aligned}
P_{i_0}(y_{l,m,n}) &= 0, \quad l = 1, ..., n, \\
P_{i_0}(z_{l,m,n}) &= 0, \quad l = 1, ..., n, \ l \neq i_0, \\
P_{i_0}(z_{i_0,m,n}) &= 1, \\
P_{i_0}^{(2j_i)}(0) &= 0, \quad i = 1, ..., m, \\
P_{i_0}^{(2j_i+1)}(0) &= 0, \quad i = 1, ..., m,
\end{aligned}$$

(with the natural modification for the case: if $z_{l,m,n} = y_{l,m,n}$ then the conditions $P_{i_0}(z_{l,m,n}) = P_{i_0}(y_{l,m,n})$ are replaced by $P_{i_0}(z_{l,m,n}) = P'_{i_0}(y_{l,m,n})$). The result now follows from the simple observation that if $f = P_{i_0}$ and $x_{l,m,n} = y_{l,m,n}$ in (1.1) we obtain

$$I_\sigma(P_{i_0}) \quad = \quad 0,$$

while if $f = P_{i_0}$ and $x_{l,m,n} = z_{l,m,n}$ in (1.1) we obtain

$$I_\sigma(P_{i_0}) \quad = \quad \lambda_{i_0,m,n}(z),$$

which is impossible, because $\lambda_{i_0,m,n}(z) > 0$, so $z_{l,m,n} = y_{l,m,n}$, $l = 1, ..., n$. On account of the uniqueness of the nodes $x_{l,m,n}$, $l = 1, ..., n$, it follows the equality of quadrature weights

$$\lambda_{l,m,n}(z) = \lambda_{l,m,n}(y),\ l = 1, ..., n,$$

and

$$\omega_{j_i,m,n}(z) = \omega_{j_i,m,n}(y),\ i = 1, ..., m.$$

This fact can be easily derived from interpolation. Thus from now on we may assume that the quadrature formula (1.1), if it exists, is unique.

II) *Existence result.* In order to establish the existence of (1.1) let us consider the space $\mathcal{S}_{2n-1,2m}$ defined by

$$\mathcal{S}_{2n-1,2m} = \left\{ P \in \mathcal{P}_{2n+2m-1};\ P^{(2j_i)}(0) = \ P^{(2j_i+1)}(0) = 0,\ i = 1, ..., m \right\}.$$

It follows from Atkinson-Sharma theorem that for any choice of t_i, $i = 1, \ldots, n$, such that $t_1 < ... < t_n$ and for all real data $\left\{ y_i^j \right\}_{i=1,j=0}^{n,1}$, there exists a unique polynomial $P \in \mathcal{S}_{2n-1,2m}$ such that

$$\begin{array}{rcll} P(t_i) & = & y_i^0, & i = 1, ..., n, \\ P'(t_i) & = & y_i^1, & i = 1, ..., n. \end{array}$$

Thus $\mathcal{S}_{2n-1,2m}$ is a space of dimension $2n$, then by the classical Krein theorem [29], there exists a unique quadrature formula of the form

$$\int_{-1}^{1} f(x)\, d\sigma(x) = \sum_{l=1}^{n} \lambda_{l,m,n}^K f(x_{l,m,n}^K) + R_{m,n}^K(f), \tag{2.1}$$

which integrates exactly all polynomials of $\mathcal{S}_{2n-1,2m}$ and such that

$$-1 < x_{1,m,n}^K < \ldots < x_{n,m,n}^K < 1,$$

and

$$\lambda_{l,m,n}^K > 0, \qquad l = 1, \ldots, n.$$

We are now ready to prove the existence of (1.1). Let $I_{2n+2m-1}(f)$ be the interpolation polynomial of $\mathcal{P}_{2n+2m-1}$ based on the information

$$\left\{ f(x_{l,m,n}^K),\ f'(x_{l,m,n}^K),\ \ l = 1, \ldots, n;\ f^{(2j_i)}(0),\ f^{(2j_i+1)}(0),\ \ i = 1, \ldots, m \right\}.$$

Then, it is well-known that

$$I_{2n+2m-1}(f)(x) = \sum_{j=0}^{1} \sum_{l=1}^{n} f^{(j)}(x_{l,m,n}^K) h_{j,l,m,n}(x) + \sum_{i=1}^{m} \sum_{k=2j_i}^{2j_i+1} f^{(k)}(0) p_{k,i,m,n}(x),$$

where $h_{j,l,m,n}$, $p_{k,i,m,n}$ are the so-called fundamental polynomial functions. When $f(x)$ is a polynomial of degree $\leq 2n+2m-1$, the interpolation polynomial coincides with $f(x)$. Thus

$$f(x)=\sum_{j=0}^{1}\sum_{l=1}^{n} f^{(j)}(x_{l,m,n}^{K})h_{j,l,m,n}(x)+\sum_{i=1}^{m}\sum_{k=2j_i}^{2j_i+1} f^{(k)}(0)p_{k,i,m,n}(x),$$

when $f\in\mathcal{P}_{2n+2m-1}$. It follows immediately that an integration of $I_{2n+2m-1}(f)$ leads to the following quadrature formula,

(2.2)

$$\int_{-1}^{1} f(x)\,d\sigma(x)=\sum_{j=0}^{1}\sum_{l=1}^{n}\lambda_{l,m,n}^{(j)}f^{(j)}(x_{l,m,n}^{K})+\sum_{i=1}^{m}\sum_{k=2j_i}^{2j_i+1}\omega_{i,m,n}^{(k)}f^{(k)}(0),\forall f\in\mathcal{P}_{2n+2m-1},$$

where

$$\lambda_{l,m,n}^{(j)}=\int_{-1}^{1} h_{j,l,m,n}(x)\,d\sigma(x),$$

and

$$\omega_{i,m,n}^{(k)}=\int_{-1}^{1} p_{k,i,m,n}(x)\,d\sigma(x). \tag{2.3}$$

Note that $h_{1,l,m,n}\in\mathcal{S}_{2n-1,2m}$, for all $l=1,...,n$, and vanishes at the nodes of (2.1), then applying (2.1) to $h_{1,l,m,n}$, we obtain $\lambda_{l,m,n}^{(1)}=0$, $l=1,...,n$. We have also $\lambda_{l,m,n}^{(0)}=\lambda_{l,m,n}^{K}$, $l=1,...,n$, this fact follows immediately by applying the quadrature formulae (2.1) and (2.2) to $h_{0,l,m,n}$, $l=1,...,n$. Hence, all weights of the free nodes in (2.2) are positive.

Now, it only remains to prove that $\omega_{i,m,n}^{(2j_i+1)}=0$, $i=1,...,m$. For this we use (2.3), the symmetry of $d\sigma$ and the fact that $p_{2j_i+1,i,m,n}$, $i=1,...,m$, are odd functions. This shows that (2.2) is a quadrature formula of the unique one (1.1) and thereby establishes the existence and uniqueness of (1.1). □

To be useful in practice, it is particularly desirable that an excellent quadrature formula should have the interlacing property. The latter is, for example, shared by all the Gaussian quadrature. Our next theorem shows that for the formula (1.1) this holds true.

Theorem 2. *Given a nonnegative even measure $d\sigma$ and nonnegative integers n,m, j_1, $j_2,...,j_m$ with $n\geq 1$. Assume that*

$$\int_{-1}^{1} f(x)\,d\sigma(x)=\sum_{i=1}^{m}\omega_{j_i,m,n}f^{(2j_i)}(0)+\sum_{l=1}^{n}\lambda_{l,m,n}f(x_{l,m,n}),\forall f\in\mathcal{P}_{2n+2m-1}, \tag{2.4}$$

and

$$\int_{-1}^{1} f(x)\,d\sigma(x)=\sum_{i=1}^{m}\omega_{j_i,m,n+1}f^{(2j_i)}(0)+\sum_{l=1}^{n+1}\lambda_{l,m,n+1}f(x_{l,m,n+1}),\forall f\in\mathcal{P}_{2n+2m+1},$$

with $-1<x_{1,m,n}<\ldots<x_{n,m,n}<1$ and $-1<x_{1,m,n+1}<\ldots<x_{n+1,m,n+1}<1$. Then the following interlacing property holds:

$$-1<x_{1,m,n+1}<x_{1,m,n}<...<x_{n,m,n}<x_{n+1,m,n+1}<1.$$

Proof. Arguing by contradiction, we suppose there exists μ with

$$x_{\mu,m,n} \notin (x_{\mu,m,n+1}, x_{\mu+1,m,n+1}) .$$

Define the polynomial $P_\mu \in P_{2n+2m-1}$ by the interpolation conditions

$$\begin{array}{lcll} P_\mu(x_{l,m,n}) & = & 0, & l = 1, ..., n, \\ P_\mu(x_{l,m,n+1}) & = & 0, & l = 1, ..., n+1,\ l \neq \mu \text{ and } l \neq \mu+1, \\ P_\mu(x_{\mu,m,n+1}) & = & 1, & \\ P_\mu^{(2j_i)}(0) & = & 0, & i = 1, ..., m, \\ P_\mu^{(2j_i+1)}(0) & = & 0, & i = 1, ..., m. \end{array}$$

(As in the proof of Theorem 1, the modifications necessary are obvious if some nodes $x_{l,m,n}$ and $x_{l,m,n+1}$ coincide.) The existence of P_μ follows from the theorem of Atkinson-Sharma once more. Taking into account that the total number of zeros of P_μ on $[-1, 1]$ does not exceed $2n+2m-1$ it easily follows that $P_\mu(x_{\mu+1,m,n+1}) > 0$. Therefore,

$$\int_{-1}^{1} P_\mu(x)\, d\sigma(x) = 0 \tag{2.5}$$

by (2.4). But, by the second quadrature formula we have

$$\int_{-1}^{1} P_\mu(x)\, d\sigma(x) = \lambda_{\mu,m,n+1} P_\mu(x_{\mu,m,n+1}) + \lambda_{\mu+1,m,n+1} P_\mu(x_{\mu+1,m,n+1}).$$

We remark that (2.5) cannot hold because

$$\lambda_{\mu,m,n+1} P_\mu(x_{\mu,m,n+1}) + \lambda_{\mu+1,m,n+1} P_\mu(x_{\mu+1,m,n+1}) > 0,$$

which gives a contradiction. This shows the required interlacing property. □

3. Characterization Theorems

The main result of this section shows that the free nodes of (1.1) are zeros of a certain orthogonal polynomial with respect to a linear functional $I_{q_{n,r}}$. We also establish that the weights relative to the interior nodes are proportional to the squares of the first components of the orthonormal eigenvectors, of the Jacobi matrix of order n associated with $I_{q_{n,r}}$. Then, we identify the orthogonal polynomial with respect to $I_{q_{n,r}}$ as a quasi-orthogonal polynomial which can be represented as characteristic polynomial of a symmetric tridiagonal matrix, this is the key to our analysis. Thus, the interior nodes and weights of (1.1) can then be computed directly by standard software for Gaussian quadrature formulae. This will facilitate considerably the computation of the functional $Q_{\sigma,m,n}$.

In order to formulate the problem precisely, we first give some addtional notation and formal definitions as well as some known results. A further particular measure, which will play an essential role, is

$$d\widehat{\sigma}(x) = x^{2j_m+2} d\sigma(x). \tag{3.1}$$

We also introduce the following quadrature formula,

$$\hat{Q}_{d\widehat{\sigma},m,n}(f; j_1, ..., j_m) = \sum_{l=1}^{n} x_{l,m,n}^{2j_m+2} \lambda_{l,m,n} f(x_{l,m,n}), \tag{3.2}$$

in which $x_{l,m,n}$, $\lambda_{l,m,n}$, $l = 1, ..., n$, are respectively the interior nodes and their respective weights in (1.1). Notice the new quadrature formula satisfies

$$\hat{Q}_{d\widehat{\sigma},m,n}(f; j_1, ..., j_m) = Q_{\sigma,m,n}(x^{2j_m+2} f; j_1, ..., j_m).$$

Thus, $\hat{Q}_{d\widehat{\sigma},m,n}$ is positive (all its weights are positive) and it integrates exactly all polynomials from $\mathcal{P}_{2n-r-1}$, where, for notational convenience, we denoted by r the following integer

$$r = 2j_m - 2m + 2. \tag{3.3}$$

Here, as well as in the remaining of this paper, we assume that such a integer has been given which satisfies

$$2 \leq r < n. \tag{3.4}$$

Let now the "nodes polynomial", which we write, for future convenience, $q_{n,r}$,

$$q_{n,r}(x) = \prod_{l=1}^{n} (x - x_{l,m,n}) \tag{3.5}$$

having as zeros the interior nodes $x_{l,m,n}$, $l = 1, ..., n$, of (1.1).

Let $f \in \mathcal{P}_{2n-1}$, dividing f by the polynomial $q_{n,r}$ gives

$$f(x) = q(x) q_{n,r}(x) + r_{q_{n,r}}(x), \tag{3.6}$$

where both q and $r_{q_{n,r}}$ are unique polynomials of degree less than or equal $n - 1$. We follow the authors [40], [46] and define a linear functional associated with $q_{n,r}$ by:

$$\begin{array}{rccl} I_{q_{n,r}}: & \mathcal{P}_{2n-1} & \to & \mathbb{R} \\ & f & \to & I_{q_{n,r}}(f) = \int_{-1}^{1} r_{q_{n,r}}(x) d\widehat{\sigma}(x). \end{array} \tag{3.7}$$

Note that the form $I_{q_{n,r}}$ depends on $d\widehat{\sigma}$, but for convenience we have suppressed this dependence to simplify the notation. This functional is positive on $\mathcal{P}_{2n-1}$ in the following sense: for every nonnegative polynomial $f \in \mathcal{P}_{2n-1}$, we have

$$I_{q_{n,r}}(f) \geq 0,$$

and

$$I_{q_{n,r}}(f) = 0, \text{ only if } f = 0. \tag{3.8}$$

The proof of this result is essentially the same as the proof in [40, p. 397]. We include it here for the sake of completeness. It is easily shown, since the quadrature formula (3.2) is exact for all elements of $\mathcal{P}_{n-1}$ ($\subset \mathcal{P}_{2n-r-1}$), that the functional $I_{q_{n,r}}$ satisfies

$$\begin{array}{rcl} I_{q_{n,r}}(f) & = & \int_{-1}^{1} r_{q_{n,r}}(x) d\widehat{\sigma}(x), \\ & = & \hat{Q}_{d\widehat{\sigma},m,n}(r_{q_{n,r}}; j_1, ..., j_m), \end{array}$$

in view of $r_{q_{n,r}}(x_{l,m,n}) = f(x_{l,m,n})$, $l = 1, ..., n$. So, we have

$$I_{q_{n,r}}(f) = \hat{Q}_{d\widehat{\sigma},m,n}(f; j_1, ..., j_m).$$

Thus, since $\hat{Q}_{d\widehat{\sigma},m,n}$ is positive (all its coefficients are positive), we deduce that $I_{q_{n,r}}(f) \geq 0$ for all nonnegative polynomial f of $\mathcal{P}_{2n-1}$, and it remains to show that the only solution of $I_{q_{n,r}}(f) = 0$ is the zero solution. This is done by applying Lukács representation theorem for nonnegative polynomials (see Szegö [42, Theorem 1.2.2, p. 4]), which ensures that if $f \in \mathcal{P}_{2n-1}$ and nonnegative in $[-1, 1]$, then f can be represented in the form

$$f(x) = (1+x)C^2(x) + (1-x)D^2(x)$$

for some polynomials C and $D \in \mathcal{P}_{n-1}$. As such and if it is non-trivial polynomial, then it has the property that it cannot vanish at *all* the nodes $x_{l,m,n}$, $l = 1, ..., n$.

The positivity of the linear functional $I_{q_{n,r}}$ (3.7) on $\mathcal{P}_{2n-1}$ gives rise to the concept of orthogonality with respect to the functional $I_{q_{n,r}}$: Two polynomials u and v are orthogonal with respect to $I_{q_{n,r}}$ if $I_{q_{n,r}}(uv) = 0$ (see [10, Chapter 1, Section 2]). In particular, there exists unique monic polynomials $\pi_n(.) = \pi_n(., I_{q_{n,r}})$ of degree n orthogonal with respect to the functional $I_{q_{n,r}}$ to all polynomials of lower degree. These orthogonal polynomials can conveniently be represented by the tridiagonal Jacobi matrix $\mathbf{J}_n(I_{q_{n,r}})$ of coefficients of the recurrence relation which they satisfy.

We begin by showing an essential property of the interior nodes of (1.1), which we record for later use.

Theorem 3. *Suppose r, $q_{n,r}$ and $I_{q_{n,r}}$ are given respectively as in (3.3), (3.5) and (3.7). Suppose further that $x_{1,m,n}, ..., x_{n,m,n}$ are n-points on the interval $(-1, 1)$, such that $x_{i,m,n} \neq x_{j,m,n}$ for all $i \neq j$. Then the n nodes $x_{1,m,n}, ..., x_{n,m,n}$ are the free nodes of the quadrature formula (1.1) if and only if they are precisely roots of the monic orthogonal polynomial π_n.*

Proof. We prove this result by using some elementary argument and showing that $\pi_n(., I_{q_{n,r}})$ is in fact $q_{n,r}$. To show this, observe that

$$\begin{aligned} I_{q_{n,r}}(q_{n,r}p) &= \hat{Q}_{d\widehat{\sigma},m,n}(q_{n,r}p; j_1, ..., j_m), \quad \text{for all } p \in \mathcal{P}_{n-1} \\ &= 0. \end{aligned} \tag{3.9}$$

Thus, since $\pi_n(., I_{q_{n,r}})$ (assumed monic of degree n) is uniquely determined by (3.9), this identifies $\pi_n(., I_{q_{n,r}})$ as $q_{n,r}$. The rest of proof proceeds exactly as usual. □

The following interesting corollary, which is obvious from Theorem 3, gives a simple and an important characterization of the free nodes in (1.1) that are recognized as eigenvalues of a symmetric tridiagonal matrix. We will use this fact later to efficiently compute them.

Corollary 1. *Suppose r, $q_{n,r}$ and $I_{q_{n,r}}$ are given respectively as in (3.3), (3.5) and (3.7). Then the n nodes $x_{1,m,n}, ..., x_{n,m,n}$ are free nodes of the quadrature formula (1.1) if and only if they are the eigenvalue of the Jacobi matrix $\mathbf{J}_n(I_{q_{n,r}})$ which defines the orthogonal polynomial with respect to the functional $I_{q_{n,r}}$.*

Let us now suppose that the interior n-nodes of (1.1) are be found as eigenvalues of the symmetric tridiagonal matrix $\mathbf{J}_n(I_{q_{n,r}})$ of Corollary 1. We now establish the following result, which simply says that the weights $\lambda_{l,m,n}$, $l = 1, ..., n$, of (1.1) are proportional to the squares of the first components of the normalized eigenvectors of $\mathbf{J}_n(I_{q_{n,r}})$. This result will later be the key to efficient computation of $\lambda_{l,m,n}$, $l = 1, ..., n$, of (1.1).

Theorem 4. *Suppose the symmetric tridiagonal matrix* $\mathbf{J}_n(I_{q_{n,r}})$ *is given as in Corollary 1. Suppose the eigenvectors of* $\mathbf{J}_n(I_{q_{n,r}})$ *are calculated such that*

$$\mathbf{J}_n(I_{q_{n,r}})\mathbf{V}_l = x_{l,m,n}\mathbf{V}_l, \qquad l = 1, ..., n,$$

with $\mathbf{V}_l^T\mathbf{V}_l = 1$ *and* $\mathbf{V}_l^T = (v_{1,l}, ..., v_{n,l})$. *Then, the weights* $\lambda_{l,m,n}$, $l = 1, ..., n$, *can be expressible in terms of the first components* $v_{1,l}$ *of* $\mathbf{V}_l$ *by*

$$\lambda_{l,m,n} = \frac{v_{1,l}^2}{x_{l,m,n}^{2j_m+2}} \int_{-1}^{1} d\widehat{\sigma}(x), \ l = 1, ..., n, \ \textit{(}l \textit{ is such that } x_{l,m,n} \neq 0\textit{)}. \tag{3.10}$$

Proof. The proof of our result is facilitated by showing that the quadrature formula (3.2) is precisely the Gaussian one for the functional $I_{q_{n,r}}$ (3.7). Indeed, this fact can be proved, if one observes, in view of the exactness of (3.2) on $\mathcal{P}_{n-1}$ ($\subset \mathcal{P}_{2n-r-1}$), that for all polynomial of degree $\leq n-1$

$$\begin{array}{rcl} I_{q_{n,r}}(f) & = & \int_{-1}^{1} f(x)d\widehat{\sigma}(x), \\ & = & \hat{Q}_{d\widehat{\sigma},m,n}(f; j_1, ..., j_m). \end{array} \tag{3.11}$$

Hence, the weights of $\hat{Q}_{d\widehat{\sigma},m,n}$ and those of the Gaussian quadrature formula (relative to $I_{q_{n,r}}$) are then identical. This result is a trivial consequence of the well known fact that the weights of a quadrature formula, which is exact on $\mathcal{P}_{n-1}$ and uses n distinct nodes, is uniquely determined. Consequently, the weights can be obtained uniquely from the linear system

$$I_{q_{n,r}}(f) = \hat{Q}_{d\widehat{\sigma},m,n}(f; j_1, ..., j_m), \text{ all } f \in \mathcal{P}_{n-1}. \tag{3.12}$$

A number of methods are known for their construction. However, the equations (3.12) do not lend to constructive purposes because of the well-known ill-conditioning associated with power moments, which is highly inconvenient for practical use. In order to combat the high condition number, it is more effective to reduce this problem to a eigenvalue problem with respect to the functional $I_{q_{n,r}}$.

By the Gram-Schmidt orthogonalization process, the positivity of the linear functional $I_{q_{n,r}}$ (3.7) on $\mathcal{P}_{2n-1}$ implies the existence of a unique system of polynomials $\left\{q_k(.) = q_k(., I_{q_{n,r}}), k = 0, ..., n\right\}$, q_k being of exact degree k, that are orthonormal with respect to $I_{q_{n,r}}$. We observe that $q_n = q_{n,r}/\sqrt{\int_{-1}^{1} q_{n,r}^2(x)d\widehat{\sigma}(x)}$ and, since $0 \leq r < n$, an easy calculation shows that we also have $q_0 = 1/\sqrt{\int_{-1}^{1} d\widehat{\sigma}(x)}$.

The remainder of the proof may be obtained by using a trivial modification of the argument for the simplest common case of Golub and Welsch [22, p. 223]. Indeed, it follows, from the fact that q_k are orthonormal polynomials for $I_{q_{n,r}}$ and from the well-known result of the zeros and the Christoffel-Darbout formula satisfied by all orthogonal polynomials, that if l is such that $x_{l,m,n} \neq 0$ then the weights $\lambda_{l,m,n}$ can be expressed in terms of $q_0, ..., q_{n-1}$, by

$$x_{l,m,n}^{2j_m+2}\lambda_{l,m,n} = \frac{1}{\sum_{k=0}^{n-1}\left[q_k(x_{l,m,n})\right]^2}, \ l = 1, ..., n. \tag{3.13}$$

Finally, to complete the proof of Theorem 4, we use the fact that $\sum_{k=0}^{n-1} v_{k,l}^2 = 1$, to prove that $v_{1,l}^2 = \lambda_{l,m,n}\left[q_0(x_{l,m,n})\right]^2$. This establishes the theorem. □

We remark that, since $\hat{Q}_{d\widehat{\sigma},m,n}(.;j_1,...,j_m)$ is a Gaussian quadrature formula for the functional $I_{q_{n,r}}$, the positivity of the weights $\lambda_{l,m,n}$, $l=1,...,n$, and the interlacing property of the nodes $x_{l,m,n}$, $l=1,...,n$, proved in Section 2, are thereby guaranteed at once, because these properties are shared by *all* Gaussian quadrature formulae.

Remark 2. *We do not know the functional $I_{q_{n,r}}$, but we do know a lot about it: We defer until Section 4 other aspects of its numerical implementation, such as the explicit or numerical construction of the coefficients of the three-term recurrence (or equivalently the Jacobi matrice) which define the polynomials orthogonal with respect to $I_{q_{n,r}}$. Note that, from an implementation standpoint, an important practical aspect of our characterization of the weights is that the expression given in (3.10) depend only on $\mathbf{J}_n(I_{q_{n,r}})$ and $d\widehat{\sigma}$. We remark also that the elegant relationships (3.10) can easily be extended to the case of an arbitrary positive quadrature formula.*

Following the same path as in [17] and [27], the next step in our quest for the quadrature formula (1.1) identifies the quasi-orthogonal polynomial (3.5). For this purpose, we first introduce the space $\mathcal{K}_{n+2j_m+1}$, which is useful subsequently, by

(3.14)

$$\mathcal{K}_{n+2j_m+1}=\left\{P=q_{n,r}R\in\mathcal{P}_{n+2j_m+1};\ P^{(2j_i)}(0)=\ P^{(2j_i+1)}(0),\ i=1,...,m\right\}.$$

Again, using Atkinson-Sharma theorem, it is easy to verify that the dimension of $\mathcal{K}_{n+2j_m+1}$ is r. Consequently, we may construct a basis $\{\Psi_0,...,\Psi_{r-1}\}$ for $\mathcal{K}_{n+2j_m+1}$ of the form

$$\Psi_i(x) \quad = \quad x^{n+2m+i}+R_i(x), \qquad i=0,...,r-1, \tag{3.15}$$

with R_i belonging to $\mathcal{P}_{n+2m-1}$. In order to define a basis of $\mathcal{P}_{2n+2m-1}$, which will play an essential role for the proof of Theorem 5, we introduce the following polynomials

$$\Phi_k(x)=x^{2m+r+k}q_{n,r}(x),\ k=0,...,n-r-1. \tag{3.16}$$

Observe that these polynomials satisfy the following orthogonality relations,

$$\int_{-1}^{1}\Phi_k(x)\ d\sigma(x)=0,\ k=0,...,n-r-1. \tag{3.17}$$

Note also that we have for, $i=1,...,m$,

$$\Phi_k^{(2j_i)}(0)=\Phi_k^{(2j_i+1)}(0)=0,\ k=0,...,n-r-1. \tag{3.18}$$

The relations (3.17) follow immediately from the quasi-orthogonality of $q_{n,r}$ (see (3.20)) and the fact that

$$\int_{-1}^{1}\Phi_k(x)\ d\sigma(x) \quad = \quad \int_{-1}^{1}x^k q_{n,r}(x)d\widehat{\sigma}(x),\ k=0,...,n-r-1.$$

Thus, it can be easily proved that the following collection of polynomials

$$\left\{x^i,\ i=0,...,n+2m-1;\ \Psi_j,\ j=0,...,r-1;\ \Phi_k,\ k=0,...,n-r-1\right\}, \tag{3.19}$$

forms a basis for $\mathcal{P}_{2n+2m-1}$. In the proof of the next theorem we shall have a use for this observation.

We are now in position to present our main result of this section, which is undoubtedly the key point to develop an algorithm for the computation of the functional $Q_{\sigma,m,n}$.

Theorem 5. *Given a nonnegative even measure $d\sigma$ and nonnegative integers n, m, $j_1,\ j_2, ..., j_m$ with $n \geq 1$. Suppose r and the functions $d\widehat{\sigma}, \Psi_0, ..., \Psi_{r-1}$ are given respectively as in (3.3), (3.1) and (3.19). Suppose further that $x_{1,m,n}, ..., x_{n,m,n}$ are n-points located symmetrically in the interval $(-1, 1)$, such that $x_{i,m,n} \neq x_{j,m,n}$ for all $i \neq j$. Then the n nodes $x_{1,m,n}, ..., x_{n,m,n}$ are the free nodes of the quadrature formula (1.1) if and only if they are zeros of a quasi-orthogonal polynomial $q_{n,r}$ of degree n and order r with respect to $d\widehat{\sigma}$ $(= x^{2j_m+2}d\sigma)$, i. e.*

$$\int_{-1}^{1} q_{n,r}(x)p(x)d\widehat{\sigma}(x) = 0, \text{ for all } p \in \mathcal{P}_{n-r-1}, \tag{3.20}$$

such that

$$\int_{-1}^{1} \Psi_i(x)d\sigma(x) = 0, \qquad i = 0, ..., r-1. \tag{3.21}$$

Proof. For the sake of notational simplicity the dependence of $Q_{d\widehat{\sigma},m,n}(f; j_1, ..., j_m)$ on $j_1, ..., j_m$, will from on be suppressed, therefore we denote

$$Q_{d\widehat{\sigma},m,n}(f; j_1, ..., j_m) = Q_{d\widehat{\sigma},m,n}(f),$$

where $Q_{d\widehat{\sigma},m,n}$ is the quadrature formula defined in (3.2). We recall that $\hat{Q}_{d\widehat{\sigma},m,n}$ integrates exactly all polynomials from $\mathcal{P}_{2n-r-1}$.

Necessity. Assume that the nodes $x_{1,m,n}, ..., x_{n,m,n}$, are those of the quadrature formula (1.1). The first part of the result follows directly from the fact that the quadrature formula (1.1) is exact on $\mathcal{P}_{2n-r-1}$. In fact, suppose that p has degree less that or equal $n-r-1$. Since $q_{n,r}(x)p(x)$ is of degree at most $2n-r-1$ and the quadrature formula (3.2) is exact for elements of $\mathcal{P}_{2n-r-1}$, we have

$$\begin{aligned} \int_{-1}^{1} q_{n,r}(x)p(x)\ d\widehat{\sigma}(x) &= \hat{Q}_{d\widehat{\sigma},m,n}(q_{n,r}p), \\ &= 0, \end{aligned}$$

therefore, $q_{n,r}$ is a quasi-orthogonal polynomial of degree n and order r with respect to the measure $d\widehat{\sigma}$.

For the second result, note first that, since $r \leq n$ then for all $i = 0, ..., r-1$, Ψ_i is a polynomial of degree $\leq 2n + 2m - 1$. Thus, the exactness of $Q_{\sigma,m,n}$ on $\mathcal{P}_{2n+2m-1}$ gives for all $i = 0, ..., r-1$,

$$\begin{aligned} \int_{-1}^{1} \Psi_i(x)d\sigma(x) &= Q_{\sigma,m,n}(\Psi_i), \\ &= 0. \end{aligned}$$

Which proves the necessity of the condition asserted in Theorem 5.

Sufficiency. Assume that (3.21) holds and that the polynomial $q_{n,r}$ is a quasi-orthogonal polynomials of degree n and order r having n distinct zeros $x_{1,m,n}, ...,$ $x_{n,m,n}$ on $(-1, 1)$. For a given function f on $(-1, 1)$, we denote by $\mathcal{I}_{n+2m-1}(f; .)$, $\{h_{l,m,n}, p_{2j_i,i,m,n}, p_{2j_i+1,i,m,n}\}$ respectively the interpolation polynomial and the set of nodal basis functions of $\mathcal{P}_{n+2m-1}$ with respect to the data

$$\left\{ f(x_{l,m,n}),\ \ l = 1, \ldots, n;\ f^{(2j_i)}(0),\ f^{(2j_i+1)}(0),\ \ i = 1, \ldots, m \right\}.$$

The existence of $\mathcal{I}_{n+2m-1}(f;.)$ and $\{h_{l,m,n}, p_{2j_i,i,m,n}, p_{2j_i+1,i,m,n}\}$ follows at once from Atkinson-Sharma theorem. Then, we have

$$I_{n+k-1}(f)(x) = \sum_{l=1}^{n} f(x_{l,m,n}) h_{l,m,n}(x) + \sum_{i=1}^{m} \sum_{k=2j_i}^{2j_i+1} f^{(k)}(0) p_{k,i,m,n}(x).$$

Since $\mathcal{I}_{n+2m-1}(f;x) = f(x)$ for all $f \in \mathcal{P}_{n+2m-1}$, we obtain the following quadrature formula

(3.22)

$$\int_{-1}^{1} f(x)\, d\sigma(x) = \sum_{l=1}^{n} \lambda_{l,m,n} f(x_{l,m,n}) + \sum_{i=1}^{m} \sum_{k=2j_i}^{2j_i+1} \omega_{i,m,n}^{(k)} f^{(k)}(0), \forall f \in \mathcal{P}_{n+2m-1},$$

where

$$\lambda_{l,m,n} = \int_{-1}^{1} h_{l,m,n}(x)\, d\sigma(x),$$

and

$$\omega_{i,m,n}^{(k)} = \int_{-1}^{1} p_{k,i,m,n}(x)\, d\sigma(x). \tag{3.23}$$

To complete the proof of the theorem it remains to show that (3.22) integrates exactly, all polynomials of degree less than or equal to $2n + 2m - 1$. This is easily accomplished by using (3.19) that every polynomial P from $\mathcal{P}_{2n+2m-1}$ can be expressed uniquely as

$$P(x) = \sum_{k=0}^{n-r-1} a_k \Phi_k(x) + \sum_{j=0}^{r-1} b_j \Psi_j(x) + R(x), \tag{3.24}$$

where R is a polynomial of degree less than $n + 2m - 1$. The integration of (7) leads to,

$$\begin{array}{rcl} \int_{-1}^{1} P(x)\ d\sigma(x) & = & \sum_{k=0}^{n-r-1} a_k \int_{-1}^{1} \Phi_k(x)\ d\sigma x) + \sum_{j=0}^{r-1} b_j \int_{-1}^{1} \Psi_j(x)\ d\sigma(x) \\ & & + \int_{-1}^{1} R(x)\ d\sigma(x). \end{array} \tag{3.25}$$

As for all $k = 0, ..., n - r + 1$, and $j = 0, ..., r - 1$, we have from (3.20) and (3.21)

$$\int_{-1}^{1} \Phi_k(x)\ d\sigma(x) = 0 \text{ and } \int_{-1}^{1} \Psi_j(x)\ d\sigma(x) = 0.$$

Therefore, by (3.22), we have

$$\int_{-1}^{1} P(x)\ d\sigma(x) = \int_{-1}^{1} R(x)\ d\sigma(x). \tag{3.26}$$

Since R is a polynomials of degree at most $n + 2m - 1$, equations (3.22) and (3.22) imply that

$$\int_{-1}^{1} P(x)\ d\sigma(x) = \sum_{l=1}^{n} \lambda_{l,m,n} R(x_{l,m,n}) + \sum_{i=1}^{m} \sum_{k=2j_i}^{2j_i+1} \omega_{i,m,n}^{(k)} R^{(k)}(0). \tag{3.27}$$

However, for $l = 1, ..., n$, and $i = 1, ..., m$, we have

$$\begin{array}{rcccccl} \Psi_j(x_{l,m,n}) & = & \Psi_j^{(2j_i)}(0) & = & \Psi_j^{(2j_i+1)}(0) & = & 0, \quad j = 0, ..., r-1, \\ \Phi_k(x_{l,m,n}) & = & \Phi_k^{(2j_i)}(0) & = & \Phi_k^{(2j_i+1)}(0) & = & 0, \quad k = 0, ..., n-r-1, \end{array}$$

which gives for all $l = 1, ..., n$, and $i = 1, ..., m$,

$$\begin{aligned} P(x_{l,m,n}) &= R(x_{l,m,n}), \\ P^{(2j_i)}(0) &= R^{(2j_i)}(0), \\ P^{(2j_i+1)}(0) &= R^{(2j_i+1)}(0). \end{aligned}$$

This implies that the quadrature formula (3.22) is exact for P, since

$$\int_{-1}^{1} P(x)\, d\sigma(x) = \sum_{l=1}^{n} \lambda_{l,m,n} P(x_{l,m,n}) + \sum_{i=1}^{m} \sum_{k=2j_i}^{2j_i+1} \omega_{i,m,n}^{(k)} P^{(k)}(0). \tag{3.28}$$

Furthermore, since $d\sigma$ is symmetric, and the support of $d\sigma$ is symmetric with respect the origin, it follows easily from uniqueness that $p_{2j_i+1,i,m,n}$, $i = 1, ..., m$, are odd polynomials. So that $\omega_{i,m,n}^{(2j_i+1)} = 0$, $i = 1, ..., m$, hold trivially. This shows that (3.28) is a quadrature formula of the form (1.1). By Theorem 1, the quadrature formula (1.1) is unique, so the nodes $x_{l,m,n}$, $l = 1, ..., n$, are indeed those of (1.1). This completes the proof of Theorem 5. □

We recall that, since $d\sigma$ is assumed an even function and its support is symmetric to the origin on $[-1, 1]$, it follows from uniqueness that the nodes polynomial $q_{n,r}$ satisfy the following property:

$$q_{n,r}(-x) = (-1)^n q_{n,r}(x). \tag{3.29}$$

Thus, $q_{n,r}$ is an odd or even function depending on whether n is even or odd. In the discussion which follows, we suppose $n = 2k + 1$ and we denote by $\widehat{\pi}_i$ the monic orthogonal polynomial of exact degree i, that are orthogonal with respect to $d\widehat{\sigma}$. Since $q_{n,r}$ is a linear combination of $\widehat{\pi}_0, \widehat{\pi}_1, ..., \widehat{\pi}_n$, the orthogonality relations of the polynomials $\{\widehat{\pi}_l\}$ and the symmetry property (3.29) yield that $q_{n,r}$ can be expressed uniquely as

$$q_{n,r}(x) = \sum_{i=0}^{k} \rho_i \widehat{\pi}_{2i+1}, (\rho_{2k+1} = 1).$$

So that, if the number of nodes n is odd then 0 is a zero of the latter. Thus, a separate treatment is required for the case n odd, because in this case it is preferable to use the following characterization of the polynomial $q_{n,r}(x)/x$, which can be useful in practice.

Theorem 6. *Given a nonnegative even measure $d\sigma$ and nonnegative integers n, m, j_1, $j_2, ..., j_m$ with $n = 2k + 1$ $(k \geq 1)$. Suppose r, $q_{n,r}$ and the measure $d\widehat{\sigma}$ are given respectively as in (3.3), (3.5) and (3.1). Then the polynomial $q_{n,r}(x)/x$ is a quasi-orthogonal polynomial of degree $n - 1$ and order $r - 2$ with respect to $d\widehat{\sigma}$, i.e.*

$$\int_{-1}^{1} (q_{n,r}(x)/x) p(x) d\widehat{\sigma}(x) = 0, \text{ for all } p \in \mathcal{P}_{n-r}, \tag{3.30}$$

or equivalently

$$q_{n,r}(x)/x = \sum_{i=0}^{(r-2)/2} \rho_{2k-2i} \widehat{\pi}_{2k-2i}, \ (\rho_{2k} = 1). \tag{3.31}$$

Proof. The argument is essentially the same that in Theorem 5, so we omit details. □

Utilizing Theorem 5, Theorem 6 may be restated in term of the quasi-orthogonal polynomial $q_{n,r}\,(x)/x$.

The computation of the functional $Q_{\sigma,m,n}$ may be considered as solved, once the Jacobi matrix $\mathbf{J}_n(I_{q_{n,r}})$ has been obtained. The basic idea of computing the coefficients in $\mathbf{J}_n(I_{q_{n,r}})$ is as follows:

The first step in our quest for $\mathbf{J}_n(I_{q_{n,r}})$ computes the coefficients of the quasi-orthogonal polynomial $q_{n,r}$ that represent it in the basis:

$$\{\widehat{\pi}_k(.) = \widehat{\pi}_k(., d\widehat{\sigma}),\ k = 0, ..., n\}\,.$$

Thus, we first express the nodes polynomial $q_{n,r}$ in terms of monic orthogonal polynomials $\{\widehat{\pi}_k\}_{k=0}^{n}$, therefore $q_{n,r}$ may be written in the following form

$$q_{n,r} = \widehat{\pi}_n + \widehat{\rho}_1\widehat{\pi}_{n-1} + \ldots + \widehat{\rho}_r\widehat{\pi}_{n-r}, \tag{3.32}$$

for some choice of real constants $\widehat{\rho}_1, \ldots, \widehat{\rho}_r$, since, by virtue of Theorem 5 $q_{n,r}$ is orthogonal to $\mathcal{P}_{n-r-1}$ with respect to the measure $d\widehat{\sigma}$. The coefficients $\{\widehat{\rho}_k\}_{k=1}^{r}$ can then computed as follows: We substitute (3.32) into the equations (3.21) and a simple calculation shows that this gives rise to a (nonlinear) system of r equations with the unknown coefficients $\widehat{\rho}_k$, $k = 1, ..., r$, (note that the polynomials Ψ_i are of the form $\Psi_i = q_{n,r}R_i$ for some $R_i \in \mathcal{P}_{2j_m+1}$.)

Thus finding these coefficients in $q_{n,r}$ amounts to solving a $r \times r$ system of nonlinear equations. Our strategy is then to choose the solution in a way which will assure that the corresponding polynomial (3.32) has all its zeros on $(-1, 1)$. To clarify this fact, we will discuss this more fully in section (4) for the particular case $r = 2$.

Once the underlying polynomial $q_{n,r}$ is available, we then appeal to Corollary 1 which tells us that the latter must have a symmetric tridiagonal matrix representation. In theory there are a number of ways to calculate zeros of a such polynomial. The most effective one is to compute them as eigenvalues of a symmetric tridiagonal matrix (just like the classical orthogonal polynomials), see, e. g., Goedecker [21]. This leads to great reductions in cost of quadratures. To use such a method for finding zeros of $q_{n,r}$, it is necessary to evaluate the elements of the Jacobi matrix $\mathbf{J}_n(I_{q_{n,r}})$, for which $q_{n,r}$ is its characteristic polynomial. If we knew the quasi-orthogonal polynomial in the form (3.32), we could in principle find its Jacobi matrix for small r, see [17, 46], but, because of nonlinearity, the complexity increases rapidly with r. We will see in the section that follows how to use this technique in the case $r = 2$.

Finally, in the light of the discussion following the statement of Corollary 1, and in view of Theorem 4, the weights in (1.1) can be computed very easily and efficiently once the interior nodes have been determined.

At this point in order to apply the results of this section, the matrix $\mathbf{J}_n(I_{q_{n,r}})$ must be investigated. This serves as a motivation for the work outlined in the next section, in which we will say even more on the Jacobi matrix $\mathbf{J}_n(I_{q_{n,r}})$.

4. Computation of the Jacobi matrix

The eigenvalue analysis described in Section 3 suggests to compute the roots of the $n-$th orthogonal polynomial with respect to $I_{q_{n,r}}$. As shown in Corollary 1, it can be computed via the Jacobi matrix $\mathbf{J}_n(I_{q_{n,r}})$, which we denote by

$$\mathbf{J}_n(I_{q_{n,r}}) = \begin{bmatrix} a_0 & \sqrt{b_1} & 0 & \cdots & 0 \\ \sqrt{b_1} & a_1 & \ddots & \ddots & \vdots \\ 0 & \ddots & \ddots & \sqrt{b_{n-2}} & 0 \\ \vdots & \ddots & \sqrt{b_{n-2}} & a_{n-2} & \sqrt{b_{n-1}} \\ 0 & \cdots & 0 & \sqrt{b_{n-1}} & a_{n-1} \end{bmatrix}_{n\times n}, \tag{4.1}$$

This matrix plays a very important role in what follows. In order to clarify the connection between the original Jacobi matrix $\mathbf{J}_n(d\widehat{\sigma})$ and $\mathbf{J}_n(I_{q_{n,r}})$, we first collect some auxiliary results for orthogonal polynomials $\{\widehat{\pi}_l\}$. Thus, the remainder of this section is mainly devoted to the details of computing the coefficients of $\mathbf{J}_n(I_{q_{n,r}})$ whose eigenvalues are the quadrature nodes. In so doing, we find a remarkable structure of $\mathbf{J}_n(I_{q_{n,r}})$.

The orthogonal polynomials $\{\widehat{\pi}_l\}$ with respect to $d\widehat{\sigma}$ (3.1) are important tools for us. We suppose them monic, i. e., they are normalized so that the leading coefficient of each $\widehat{\pi}_l$ is 1. Since $d\widehat{\sigma}$ is an even measure, then these orthogonal polynomials satisfy a three-term recurrence relation of the form

$$\begin{array}{rcll} \widehat{\pi}_{-1}(x) & = & 0,\ \widehat{\pi}_0(x) = 1, & \\ \widehat{\pi}_{l+1}(x) & = & x\widehat{\pi}_l(x) - \widehat{\beta}_l\widehat{\pi}_{l-1}(x), & l = 0, 1, 2, ..., \end{array} \tag{4.2}$$

where the coefficients $\left\{\widehat{\beta}_l, l = 1, ...\right\}$ are given by the well-known formulae of Stieltjes:

$$\widehat{\beta}_0 = \widehat{\beta}_0(d\widehat{\sigma}) = I_{d\widehat{\sigma}}\left(\widehat{\pi}_0^2\right),\ \widehat{\beta}_l = \widehat{\beta}_l(d\widehat{\sigma}) = \frac{I_{d\widehat{\sigma}}\left(\widehat{\pi}_l^2\right)}{I_{d\widehat{\sigma}}\left(\widehat{\pi}_{l-1}^2\right)},\ l = 1, ..., \tag{4.3}$$

with $I_{d\widehat{\sigma}}$ being the functional

$$I_{d\widehat{\sigma}}(u) = \int_{-1}^{1} u(x)d\widehat{\sigma}(x).$$

Note that $\widehat{\beta}_l > 0$. These coefficients (assumed known or computable) define an unreduced symmetric tridiagonal matrix $\mathbf{J}_n(d\widehat{\sigma})$, the so-called Jacobi matrix,

$$\mathbf{J}_n(d\widehat{\sigma}) = \begin{bmatrix} 0 & \sqrt{\widehat{\beta}_1} & 0 & \cdots & 0 \\ \sqrt{\widehat{\beta}_1} & 0 & \ddots & \ddots & \vdots \\ 0 & \ddots & \ddots & \sqrt{\widehat{\beta}_{n-2}} & 0 \\ \vdots & \ddots & \sqrt{\widehat{\beta}_{n-2}} & 0 & \sqrt{\widehat{\beta}_{n-1}} \\ 0 & \cdots & 0 & \sqrt{\widehat{\beta}_{n-1}} & 0 \end{bmatrix}_{n\times n}. \tag{4.4}$$

With the vector

$$\widehat{\mathbf{v}}_n(x) = (\widehat{\pi}_0(x), \widehat{\pi}_1(x), ..., \widehat{\pi}_n(x))^T$$

we can conveniently rewrite the three-term recurrence relation (4.2) in matrix notation

$$x\widehat{\mathbf{v}}_{n-1}(x) = \mathbf{J}_n(d\widehat{\sigma})\widehat{\mathbf{v}}_{n-1}(x) + \widehat{\beta}_n\widehat{\pi}_n(x)\mathbf{e}_n$$

where $\mathbf{e}_n = (0, 0, ..., 1)^T$ denotes the n-th unit vector.

In order to construct (1.1), we need to know the coefficients of its Jacobi matrix $\mathbf{J}_n(I_{q_{n,r}})$. To do this, we begin by proving an essential result which reduces considerably the number of unknown coefficients in $\mathbf{J}_n(I_{q_{n,r}})$. In the following the symbol $[x]$ denotes as usual the greatest integer in the real number x.

Lemma 1. *Suppose $d\widehat{\sigma}$, $\mathbf{J}_n(I_{q_{n,r}})$ and $\mathbf{J}_n(d\widehat{\sigma})$ are given respectively as in (3.1), (4.1) and (4.4). Then, for any integer r, such that $0 \leq r < n$, the following hold:*

(a) the first $n - [r/2] - 1$ diagonal elements of $\mathbf{J}_n(I_{q_{n,r}})$ are equal to 0,

and

(b) the first $n - [(r+1)/2]$ subdiagonal elements of $\mathbf{J}_n(I_{q_{n,r}})$ equal those of $J_n(d\widehat{\sigma})$.

Proof. We first note that the recursion coefficients a_k, b_k in turn are expressible in terms of the monic orthogonal polynomials q_k relative to the functional $I_{q_{n,r}}$ as

$$\begin{aligned} a_k &= \frac{I_{q_{n,r}}\left(xq_k^2\right)}{I_{q_{n,r}}\left(q_k^2\right)}, k = 0, ..., n-1, \\ b_0 = I_{q_{n,r}}\left(q_0^2\right),\ b_k &= \frac{I_{q_{n,r}}\left(q_k^2\right)}{I_{q_{n,r}}\left(q_{k-1}^2\right)}, k = 1, ..., n-1. \end{aligned} \tag{4.5}$$

On the other hand, if f is an arbitrary polynomial in $\mathcal{P}_{2n-r-1}$ then it can be written uniquely as $f = qq_{n,r} + h$, where $q \in \mathcal{P}_{n-r-1}$ and $h \in \mathcal{P}_{n-r-1}$. This implies whenever f is a polynomial of degree less than or equal to $2n - r + 1$,

$$\begin{aligned} I_{q_{n,r}}(f) &= \hat{Q}_{d\widehat{\sigma},m,n}(h; j_1, ..., j_m), \\ &= \hat{Q}_{d\widehat{\sigma},m,n}(f; j_1, ..., j_m), \end{aligned}$$

where the trivial relations $f(x_{l,m,n}) = h(x_{l,m,n}), l = 1, ..., n$, have been used in the last equality. Thus, since $\hat{Q}_{d\widehat{\sigma},m,n}$ is exact for all polynomials of degree $\leq 2n-r-1$, we conclude that

$$I_{q_{n,r}}(f) = I_{d\widehat{\sigma}}(f),\ \forall f \in \mathcal{P}_{2n-r-1}. \tag{4.6}$$

Moreover, for every $k = 0, ..., n-[(r+1)/2]$, and $f \in \mathcal{P}_{k-1}$, we have $q_k f \in \mathcal{P}_{2n-r-1}$ and therefore, by (4.6),

$$\begin{aligned} I_{q_{n,r}}(q_k f) &= I_{d\widehat{\sigma}}(q_k f), \\ &= 0. \end{aligned}$$

This means that q_k is also orthogonal to $\mathcal{P}_{k-1}$ with respect to the linear functional $I_{d\widehat{\sigma}}$. Since the two polynomials q_k and $\widehat{\pi}_k$ have the same leading coefficient, it follows from uniqueness that they must be identical. This identifies q_k as $\widehat{\pi}_k$, for $k = 0, ..., n-[(r+1)/2]$. The proof of (a) and (b) is completed by using the formulae (4.3) and (4.5). □

This Lemma shows that the first desired $2n - r - 1$ recursion coefficients in the sequence $\{a_0, b, a_1, b_1, ...\}$ equal the corresponding coefficients belonging the measure $d\widehat{\sigma}$ in the known sequence $\left\{0, \widehat{\beta}_0, 0, \widehat{\beta}_1, ...\right\}$. So, an easy calculation shows that $\mathbf{J}_n(I_{q_{n,r}})$ contains only r parameters. The basic idea of computing them is as follows. Suppose that the nodes polynomial $q_{n,r}$ has been determined as a quasi-orthogonal polynomial degree n and order r, one can therefore obtain the remaining

r unknown coefficients of $\mathbf{J}_n(I_{q_{n,r}})$ by the condition that $q_{n,r}$ is its characteristic polynomial. However, in order to determine this polynomial, we need to solve a nonlinear system of r equations with the r unknown coefficients $\widehat{\rho}_k$. Of course, this is in general not so easy for large r to evaluate such coefficients by explicit expressions. Nevertheless, as we shall see in the next section, we are able to determine them analytically at least for small values of r that are of interest in applications.

We are thereby interested in the following problem:

Problem 1. *Given a quasi-orthogonal polynomial in the form* $q_{n,r} = \widehat{\pi}_n + \widehat{\rho}_1\widehat{\pi}_{n-1} + \ldots + \widehat{\rho}_r\widehat{\pi}_{n-r}$. *As such how it can be represented as characteristic polynomial of a symmetric tridiagonal matrix.*

Explicit results in this direction, involving an arbitrary measure, indeed have already been obtained in [17] for the case $r = 1, 2, 3$ or 4. We reformulate it, in our situation, as follows:

Theorem 7. *Let* $q_{n,4}$ *be a quasi-orthogonal polynomial of the form*

$$q_{n,4} = \widehat{\pi}_n + \rho_1\widehat{\pi}_{n-1} + \ldots + \rho_4\widehat{\pi}_{n-4},$$

where Then $q_{n,4}$ *has a symmetric tridiagonal matrix representation of the form*

$$q_{n,4}(x) = \det(xI_n - J_n^*)$$

with

$$J_n^* = \begin{bmatrix} 0 & \sqrt{\widehat{\beta}_1} & 0 & 0 & \cdots & 0 \\ \sqrt{\widehat{\beta}_1} & 0 & \sqrt{\widehat{\beta}_2} & 0 & \ddots & \vdots \\ 0 & \sqrt{\widehat{\beta}_2} & \ddots & \ddots & 0 & 0 \\ 0 & 0 & \ddots & 0 & \sqrt{\widehat{\beta}_{n-2} - v_2} & 0 \\ \vdots & \ddots & 0 & \sqrt{\widehat{\beta}_{n-2} - v_2} & -u_2 & \sqrt{\widehat{\beta}_{n-1} - v_1} \\ 0 & \ldots & 0 & 0 & \sqrt{\widehat{\beta}_{n-1} - v_1} & -u_1 \end{bmatrix}_{n\times n}$$

if and only if

$$\begin{array}{rcl} \rho_4 & < & \widehat{\beta}_{n-3}\widehat{\beta}_{n-2}, \\ v_1 & \leq & \widehat{\beta}_{n-1}, \end{array} \tag{4.7}$$

and

$$\begin{array}{rcl} v_1 & = & \rho_2 - v_2 - u_2u_1, \\ u_1 & = & \rho_1 - u_2 \\ u_2 & = & \dfrac{\rho_3 - v_2\rho_1}{\widehat{\beta}_{n-2} - v_2}, \\ v_2 & = & \dfrac{\rho_4}{\widehat{\beta}_{n-3}}. \end{array}$$

The main results of the previous Theorem were given for the particular case $r \leq 4$. Of course, it could have been carried over to $r \geq 5$, but the problem is mainly computational. A more general form of this result has already appeared in [46, Theorem 4.1].

Note that, the quasi-orthogonal polynomial considered here has a symmetric tridiagonal matrix representation, then the inequalities (4.7) are automatically satisfied.

As remarked previously for quadrature formula, it is important to have all its nodes in the support of the measure. We shall see in the next section when trying to construct the quasi-orthogonal polynomials $q_{n,r}$ that there are, in general, only two polynomials which satisfy the conditions (3.21).

Then, we are faced with the following problem:

Problem 2. *How should one select the polynomial which has all its zeros in* $(-1,1)$?

In general, it is easy for small r to diagnose whether a quasi-orthogonal polynomial has all its zeros located in $(-1,1)$ once its coefficients $\{\widehat{\rho}_k\}$ are available: we refer the interested reader to the discussion of that point in [14] or [45]. For the case $r=2$, the following Theorem gives an easy way which ensures that a quasi-orthogonal polynomials has no exterior roots. This result will play particularly a key role in our discussion later in obtaining an explicit expression for the nodes polynomial of (1.1) for the particular case $r=2$.

Theorem 8. *Let* $q_{n,2}$ *be a quasi-orthogonal polynomial of degree* n*, and order* 2*, of the form*

$$q_{n,2} = \widehat{\pi}_n - \rho_2 \widehat{\pi}_{n-2}.$$

Suppose that $\rho_2 \leq \overset{\wedge}{\beta}_{n-1}$*, then* $q_{n,2}$ *has all its zeros in* $(-1,1)$ *if and only if*

$$\rho_2 \leq \min\left\{ \frac{\widehat{\pi}_n(-1)}{\widehat{\pi}_{n-2}(-1)}, \frac{\widehat{\pi}_n(1)}{\widehat{\pi}_{n-2}(1)} \right\}.$$

Proof. By Theorem 7, the n zeros $x_1 < x_2 < ... < x_n$ of $q_{n,2}$ are distinct eigenvalues of the symmetric tridiagonal matrix

$$\mathbf{J}_n(q_{n,2}) = \begin{bmatrix} 0 & \sqrt{\widehat{\beta}_1} & 0 & 0 & \cdots & 0 \\ \sqrt{\widehat{\beta}_1} & 0 & \sqrt{\widehat{\beta}_2} & 0 & \ddots & \vdots \\ 0 & \sqrt{\widehat{\beta}_2} & \ddots & \ddots & 0 & 0 \\ 0 & 0 & \ddots & 0 & \sqrt{\widehat{\beta}_{n-2}} & 0 \\ \vdots & \ddots & 0 & \sqrt{\widehat{\beta}_{n-2}} & 0 & \sqrt{\widehat{\beta}_{n-1} - \rho_2} \\ 0 & \dots & 0 & 0 & \sqrt{\widehat{\beta}_{n-1} - \rho_2} & 0 \end{bmatrix}_{n \times n}.$$

Let the $n-1$ zeros of $\widehat{\pi}_{n-1}$ be ordered according to $t_1 < t_2 < ... < t_{n-1}$. Obviously, these nodes are eigenvalues of the $(n-1)\times(n-1)$ leading submatrix of the matrix $\mathbf{J}_n(q_{n,2})$, then Cauchy's interlace theorem (see, e. g., [30, p. 186]) can be applied to give the following interlacing property of the nodes:

$$x_1 < t_1 < x_2 < ... < t_{n-1} < x_n,$$

which implies that $q_{n,2}$ has $n-2$ zeros on $(-1,1)$. Note also that, $q_{n,2}$ can have at most two exterior nodes, one at each end point, because of the interlacing property. Since the leading coefficient of $q_{n,2}$ is 1, then $q_{n,2}(1)$ is positive if only if it has no zeros greater that or equal to 1. According that $\widehat{\pi}_n(1)$ and $\widehat{\pi}_{n-2}(1)$ are both positive, and hence $x_n \in (-1,1)$ if only if $\rho_2 \leq \dfrac{\widehat{\pi}_n(1)}{\widehat{\pi}_{n-2}(1)}$; the derivation of x_1 is similar. □

5. Computation of Turán quadratures

In this section, we see how the above theory is put into practice to construct a new class of quadrature formulae of type (1.1). As with the earlier results quoted above, the amount of computation in our approach increases rapidly with r. To shorten our presentation and to better illustrate our approach we only concentrate on the devivation of our quadrature in the case $r = 2$. More precisely, we shall examine the problem of computing the following quadrature formula,

$$\int_{-1}^{1} f(x)d\sigma(x) \approx \omega_{2,n} f''(0) + \sum_{l=1}^{n} \lambda_{l,n} f(x_{l,n}), \tag{5.1}$$

which is exact for all $f \in \mathcal{P}_{2n+1}$.

Note that, as we have previously pointed out at the introduction, the present quadrature offers another solution for the problem of Turán ([43, Problem XXXIII]).

Since the construction of (5.1) is rather technical, we sketch the main ideas. To begin, observe that Theorem 1 guarantees the existence and uniqueness of (5.1), with respect to $d\sigma$ on the interval $[-1, 1]$. Moreover, in this particular case (using the values consistent with our notations of Section 3) the relevant parameters are $m = 1$, $j_1 = 1$ and $r = 2$.

Following the analysis of the previous sections, the first step in our method is to establish some important information about the entries of $\mathbf{J}_n(I_{q_{n,2}})$ and the coefficients of the quasi-orthogonal polynomial $q_{n,2}$ associated to (5.1). The lemma 1 in Section 4 furthers this aim, since, in the present case, it reduces the problem of finding the Jacobi matrix $\mathbf{J}_n(I_{q_{n,2}})$ (4.1) to compute the remaining two its unknown coefficients a_{n-1} and b_{n-1}. Indeed, here the known elements of Jacobi matrix are

$$\begin{aligned} a_i &= 0, \quad i = 0, ..., n-3 \\ b_i &= \widehat{\beta}_i, \quad i = 1, ..., n-2, \end{aligned}$$

where $\widehat{\beta}_k$ being the recurrence coefficients associated with the orthogonal polynomials

$$\left\{\widehat{\pi}_n = \widehat{\pi}_n(.; d\widehat{\sigma}(x) \ (= x^4 d\sigma(x))\right\}_{n=0,1,2,..,.}$$

This observation drastically simplifies the calculation issues associated with (5.1) which we will carry out in what follows.

We are now faced with the problem of computing the remaining unknown coefficients a_{n-1}, a_{n-2} and b_{n-1} in the Jacobi matrix $\mathbf{J}_n(I_{q_{n,2}})$. Thanks to Theorem 7, this task can be achieved by computing the quasi-orthogonal polynomial induced by (5.1). Note that, in view of Theorem 5, the latter is a quasi-orthogonal polynomial of degree n and order 2. Thus, we see that (5.1) exists if its nodes are roots of a polynomial, which can be expressed in term of $\{\widehat{\pi}_{n-k}\}_{k=0}^{2}$ as

$$q_{n,2} = \widehat{\pi}_n + c_n \widehat{\pi}_{n-1} + d_n \widehat{\pi}_{n-2}. \tag{5.2}$$

So we have two unknown coefficients in $q_{n,2}$. To find them, recall that $q_{n,2}$ is an odd or even function depending on whether n is even or odd, then it is clear that there is only one free parameter in $q_{n,2}$, since, due to symmetry, one obtains $c_n = 0$. Thus the polynomial to be studied now takes the form

$$q_{n,2} = \widehat{\pi}_n + d_n \widehat{\pi}_{n-2}. \tag{5.3}$$

Furthermore, let us note that, applying Theorem 7 allows us to infer that the unknown coefficient a_{n-2} and a_{n-1} are equal to 0. Thus, the problem of finding (5.1) is essentially solved if either the parameters d_n and b_{n-1} in $\mathbf{J}_n(I_{q_{n,2}})$ and $q_{n,2}$ respectively can be evaluated analytically or at most numerically. We remark that once d_n is know b_{n-1} can be determined uniquely by the representation Theorem 7.

Now, we proceed by first discussing how one can determine the remaining free parameters d_n in (5.3), and then outline the basic steps of a numerical algorithm for constructing (5.1). The parity of n plays a basic role in the computation of the latter, for this reason, we have to distinguish between two cases. First we begin with:

5.1. ***Case I:*** n **is odd.** This is the simplest case. The key observation is that in view of Theorem 6 the quasi-orthogonal polynomial $q_{n,2}(x)/x$ coincides with the monic orthogonal polynomial $\widehat{\pi}_{n-1}$. Thus, the nodes of (5.1) are precisely zeros of the polynomial

$$q_{n,2}(x) \quad = \quad x\widehat{\pi}_{n-1}(x). \tag{5.4}$$

Since $x\widehat{\pi}_{n-1}$ can be written (uniquely) as a linear combination of $\widehat{\pi}_0, \widehat{\pi}_1, ..., \widehat{\pi}_n$, the orthogonality relations of $\{\widehat{\pi}_l\}$ yield

$$q_{n,2}(x) = \gamma_n\widehat{\pi}_n + \gamma_{n-1}\widehat{\pi}_{n-1} + \gamma_{n-2}\widehat{\pi}_{n-2}.$$

Clearly $\gamma_n = 1$, since $q_{n,2}, \widehat{\pi}_n$ are monic polynomials having the same degree. Moreover, we also have

$$\begin{array}{rcl} \gamma_{n-1} & = & \int_{-1}^{1} x\widehat{\pi}_{n-1}^2(x)d\widehat{\sigma}(x)/\int_{-1}^{1}\widehat{\pi}_{n-1}^2(x)d\widehat{\sigma}(x), \\ & = & 0, \end{array}$$

since the integrand $x\widehat{\pi}_{n-1}^2$ and is an odd function, and so, by virtue of

$$\gamma_{n-2} = \int_{-1}^{1} x\widehat{\pi}_{n-1}(x)\widehat{\pi}_{n-2}(x)d\widehat{\sigma}(x)/\int_{-1}^{1}\widehat{\pi}_{n-2}^2(x)d\widehat{\sigma}(x),$$

the expression of $q_{n,2}$ becomes

$$q_{n,2}^o := q_{n,2} = \widehat{\pi}_n + \left(\int_{-1}^{1} x\widehat{\pi}_{n-1}(x)\widehat{\pi}_{n-2}(x)d\widehat{\sigma}(x)/\int_{-1}^{1}\widehat{\pi}_{n-2}^2(x)d\widehat{\sigma}(x)\right)\widehat{\pi}_{n-2}. \tag{5.5}$$

Thus, the remaining free parameter d_n in (5.3) is equal to

$$d_n^o := d_n = \gamma_{n-2}. \tag{5.6}$$

Having determined the polynomial $q_{n,2}$, we now debate the principal topic of this case, that of determining the nodes which occur in (5.1). So we need investigate the zeros of $q_{n,2}$. For this, we shall now discuss the determination of the remaining parameter in the Jacobi matrix $\mathbf{J}_n^o(I_{q_{n,2}}) := \mathbf{J}_n(I_{q_{n,2}})$. Our solution is based on the fact that the polynomial $q_{n,2}$ admits the following matrix representation (see Theorem 7)

$$q_{2,n}(x) = \det(x\mathbf{I}_n - \mathbf{J}_n^o(I_{q_{n,2}})),$$

where

$$\mathbf{J}_n^o(I_{q_{n,2}}) = \begin{bmatrix} 0 & \sqrt{\widehat{\beta}_1} & 0 & \dots & 0 \\ \sqrt{\widehat{\beta}_1} & 0 & \ddots & \ddots & \vdots \\ 0 & \ddots & \ddots & \sqrt{\widehat{\beta}_{n-2}} & 0 \\ \vdots & \ddots & \sqrt{\widehat{\beta}_{n-2}} & 0 & \sqrt{\widehat{\beta}_{n-1} - d_n^o} \\ 0 & \cdots & 0 & \sqrt{\widehat{\beta}_{n-1} - d_n^o} & 0 \end{bmatrix}_{n\times n}. \tag{5.7}$$

Thus, it follows from (5.7) that the unknown coefficient b_{n-1} in the Jacobi matrix $\mathbf{J}_n(I_{q_{n,2}})$ is given by

$$b_{n-1} = \widehat{\beta}_{n-1} - d_n^o. \tag{5.8}$$

Therefore, in this particular case, the nodes $x_{l,n}$ of (5.1) can be computed conveniently as the well-studied ordinary Gaussian quadrature formulae.

It remains to calculate the weights $\lambda_{l,n}$ in (5.1). To do this, we suppose that the eigenvectors of $\mathbf{J}_n^o(I_{q_{n,2}})$ are computed such that

$$\mathbf{J}_n^o(I_{q_{n,2}})\mathbf{V}_{l,n} = x_{l,n}\mathbf{V}_{l,n}, \qquad l = 1, ..., n,$$

with $\mathbf{V}_{l,n}^T\mathbf{V}_{l,n} = 1$ and $\mathbf{V}_{l,n}^T = (v_{1,l,n}, ..., v_{n,l,n})$. Then, as for the ordinary Gauss quadrature formula, it follows from Theorem 4 that the coefficients $\lambda_{l,n}$ are expressible in terms of the first components $v_{1,l,n}$ of $\mathbf{V}_{l,n}$ by

$$\lambda_{l,n} = \frac{v_{1,l,n}^2}{x_{l,n}^4}\int_{-1}^{1} d\widehat{\sigma}(x),\ l = 1, ..., n \text{ and } l \text{ is such that } x_{l,n} \neq 0.$$

We finally have that the weights $\omega_{2,n}$ and $\omega_{2,n}(0)$ (the weight relative to the node 0) in (5.1), in turn is expressible in term of the orthogonal polynomial $\widehat{\pi}_{n-1}$ as

$$\begin{aligned} \omega_{2,n} &= \frac{\int_{-1}^{1} x^2\widehat{\pi}_{n-1}(x)d\sigma(x)}{2\widehat{\pi}_{n-1}(0)}, \\ \omega_n(0) &= 2\left(\frac{\int_{-1}^{1}\widehat{\pi}_{n-1}(x)d\sigma(x)}{\int_{-1}^{1} x^2\widehat{\pi}_{n-1}(x)d\sigma(x)} - \frac{\widehat{\pi}'_{n-1}(0)}{2\widehat{\pi}_{n-1}(0)}\right)\omega_{2,n}. \end{aligned}$$

It should be noted that in the present case the nonzero nodes in (5.1) coincide with the nonzero nodes in the quadrature formula,

$$\int_{-1}^{1} f(x)d\sigma(x) = \sum_{i=0}^{3}\omega_{i,n}f^{(i)}(0) + \sum_{l=1}^{n-1}\lambda_{l,n}f(x_{l,n}), \forall f \in \mathcal{P}_{2n+1},$$

discussed by Stancu and Stroud [41, p. 387], since the coefficients of the odd derivatives $f^{(1)}(0)$ and $f^{(3)}(0)$ are equal to zero.

5.2. ***Case II:*** n **is even.** In the present case, which is treated differently, we proceed in the following manner. Our first goal is to construct a system of nonlinear equations implementing conditions (3.21) of Theorem 5 for the unknown coefficient d_n of the nodes polynomial $q_{n,2}$ (5.3). Thus, at first, we obtain from (3.14) that

$$\mathcal{K}_{n+3} = span\{\Psi_0, \Psi_1\},$$

where Ψ_0 and Ψ_1 are two polynomials of exact degree $n+3$, which satisfy the interpolation conditions

$$\begin{array}{rcccccll} \Psi_0(x_{l,n}) & = & \Psi_0''(0) & = & \Psi_0'''(0) & = & 0, & l=1,...,n, \\ \Psi_1(x_{l,n}) & = & \Psi_1''(0) & = & \Psi_1'''(0) & = & 0, & l=1,...,n. \end{array}$$

Here, we are forced to look for an appropriate base for the space $\mathcal{K}_{n+3}$ in a way which will make these conditions simple to determine the parameter d_n in $q_{n,2}$. To do so, we first note that from the definition of $\mathcal{K}_{n+3}$ its elements also vanish at the nodes of (5.1). Moreover, since n is assumed even it follows that $q_{n,2}$ is an even function, it can then be easily verified that the polynomials

$$\Psi_0(x) = (x^3 + x^2 - \frac{2q_{n,2}(0)}{q_{n,2}''(0)})q_{n,2}(x), \tag{5.9}$$

and

$$\Psi_1(x) = (x^3 - x^2 + \frac{2q_{n,2}(0)}{q_{n,2}''(0)})q_{n,2}(x), \tag{5.10}$$

form a basis for $\mathcal{K}_{n+3}$. We do not present the coefficient of the elements of $\{\Psi_0, \Psi_1\}$ in term of the standard power expansion, but in place we express them in term of the nodes polynomial $q_{n,2}$, such a representation is the most convenient form for computation, since it unveils important relationships between the expansion coefficients and the nodes polynomial $q_{n,2}$. This is the basic idea behind the construction of (5.1) in the present case.

On the other hand, according to Theorem 5, the polynomials Ψ_0 and Ψ_1 defined in (5.9) and (5.10) must also satisfy the orthogonality relations

$$\int_{-1}^{1} \Psi_0(x)d\sigma(x) = \int_{-1}^{1} \Psi_1(x)d\sigma(x) = 0.$$

Thus, by symmetry and taking into account that $q_{n,2}$ is an even function, the remaining unknown parameter d_n in (5.3) is then a solution of

$$\frac{\int_{-1}^{1} x^2 q_{n,2}(x)d\sigma(x)}{\int_{-1}^{1} q_{n,2}(x)d\sigma(x)} = \frac{2q_{n,2}(0)}{q_{n,2}''(0)}. \tag{5.11}$$

Since $q_{n,2} = \widehat{\pi}_n + d_n\widehat{\pi}_{n-2}$, using an elementary manipulation one may show that the last equation can be re-expressed in the interesting and simpler form of a quadratic polynomial,

$$a\rho^2 + b\rho + c = 0. \tag{5.12}$$

Of course, the remaining unknown coefficient d_n in $q_{n,2}$ will be certainly found amongst roots of the (5.12). Then, the next important issue that we face here is how to assess the exact value of d_n. Now, by Theorem 8 the only solution $d_n^e := d_n$ of (5.12) such that

$$d_n^e \leq \min\left\{\frac{\widehat{\pi}_n(-1)}{\widehat{\pi}_{n-2}(-1)}, \frac{\widehat{\pi}_n(1)}{\widehat{\pi}_{n-2}(1)}\right\}$$

guarantees that the polynomial,

$$q_{n,2}^e := q_{n,2} = \widehat{\pi}_n + d_n^e\widehat{\pi}_{n-2}, \tag{5.13}$$

has n distinct real zeros all located in the open interval $(-1,1)$.

We now turn to the problem of determining the Jacobi matrix $\mathbf{J}_n^e(I_{q_{n,2}}) := \mathbf{J}_n(I_{q_{n,2}})$. As before, the following construction provides a tool for finding its remaining unknown coefficient b_{n-1}, once the parameter d_n in (5.3) is obtained. We may rewrite, using Theorem 7, the polynomial $q_{n,2}$ in matrix notation as follows

$$q_{n,2}(x) = \det(x\mathbf{I}_n - \mathbf{J}_n^e(I_{q_{n,2}})),$$

where

(5.14)

$$\mathbf{J}_n^e(I_{q_{n,2}}) = \begin{bmatrix} 0 & \sqrt{\widehat{\beta}_1} & 0 & \dots & 0 \\ \sqrt{\widehat{\beta}_1} & 0 & \ddots & \ddots & \vdots \\ 0 & \ddots & \ddots & \sqrt{\widehat{\beta}_{n-2}} & 0 \\ \vdots & \ddots & \sqrt{\widehat{\beta}_{n-2}} & 0 & \sqrt{\widehat{\beta}_{n-1} - d_n^e} \\ 0 & \cdots & 0 & \sqrt{\widehat{\beta}_{n-1} - d_n^e} & 0 \end{bmatrix}_{n\times n},$$

a fact which allows us to compute the nodes $x_{l,n}$ and their respective weights $w_{l,n}$ in (5.1) by standard techniques via an eigensystem problem for the (symmetric, tridiagonal) Jacobi matrix $\mathbf{J}_n^e(I_{q_{n,2}})$.

Finally, as above, a simple calculation shows that an explicit formula for the weight $\omega_{2,n}$ in (5.1) is

$$\omega_{2,n} \quad = \quad \frac{\int_{-1}^{1} x^2 q_{n,2}^e(x) d\sigma(x)}{2q_{n,2}^e(0)}.$$

An alternative approach to constructing (5.1) when n is odd, without the use of Theorem 6, would be to proceed analogously as in the present case. Indeed, to adapt the latter to the odd case, we just have to find a more workable basis of $\mathcal{K}_{n+3}$. We can do even better by observing that, when n is odd the polynomials

$$\Psi_0(x) = (x^3 + 1)q_{n,2}(x)$$

and

$$\Psi_1(x) = (x^3 - 1)q_{n,2}(x),$$

form a basis for $\mathcal{K}_{n+3}$. Thus, in view of (3.21) and the parity of $q_{n,2}$, the resulting quasi-orthogonal polynomial $q_{n,2}$ turns out to be identical to the one given by (5.5), so the argument as in the odd case can be carried over verbatim to construct (5.1).

One particularly interesting point to note from either $\mathbf{J}_n^o(I_{q_{n,2}})$ or $\mathbf{J}_n^e(I_{q_{n,2}})$ is that their structures are unexpectedly simple. Since, in each case, to evaluate them, it suffices to take the same set of recurrence coefficients as when computing the Gauss quadrature formula $G(n; d\widehat{\sigma})$, except that the last coefficient $\widehat{\beta}_{n-1}$ is modified. The rest of computation proceeds exactly as usual. This permits a more straightforward implementation and simplifies coding.

We summarize this construction process in the following theorem.

Theorem 9. *Given a nonnegative even measure $d\sigma$ and nonnegative integers n. Define $d\widehat{\sigma} = x^4 d\sigma$. Suppose the quasi-orthogonal polynomials $q_{n,2}^o$, $q_{n,2}^e$ and the matrices $\mathbf{J}_n^o(I_{q_{n,2}})$, $\mathbf{J}_n^e(I_{q_{n,2}})$ are given respectively as in (5.5), (5.13) and (5.7),*

(5.14). Then, there exists a unique quadrature formula of the form,

$$\int_{-1}^{1} f(x)d\sigma(x) \approx \omega_{2,n} f''(0) + \sum_{l=1}^{n} \lambda_{l,n} f(x_{l,n}), \tag{5.15}$$

which integrates exactly all polynomial from $\mathcal{P}_{2n+1}$. With:

If n is odd, the nodes $x_{l,n}, l = 1, ..., n$, in (5.15) are the eigenvalues of $\mathbf{J}_n^o(I_{q_{n,2}})$ and the weights $\omega_{2,n}$ and $\omega_n(0)$ are given by

$$\begin{aligned} \omega_{2,n} &= \frac{\int_{-1}^{1} x^2 \widehat{\pi}_{n-1}(x)d\sigma(x)}{2\widehat{\pi}_{n-1}(0)}, \\ \omega_n(0) &= 2\left(\frac{\int_{-1}^{1} \widehat{\pi}_{n-1}(x)d\sigma(x)}{\int_{-1}^{1} x^2\widehat{\pi}_{n-1}(x)d\sigma(x)} - \frac{\widehat{\pi}''_{n-1}(0)}{2\widehat{\pi}_{n-1}(0)}\right)\omega_{2,n}. \end{aligned} \tag{5.16}$$

If n is even, the nodes $x_{l,n}, l = 1, ..., n$, in (5.15) are the eigenvalues of $\mathbf{J}_n^e(I_{q_{n,2}})$ and the weight $\omega_{2,n}$ is given by

$$\omega_{2,n} = \frac{\int_{-1}^{1} x^2 q^e_{n,2}(x)d\sigma(x)}{2q^e_{n,2}(0)}. \tag{5.17}$$

In all cases, the weights $\lambda_{l,n}$ are all positive and are given by

$$\lambda_{l,n} = \frac{v^2_{1,l,n}}{x^4_{l,n}} \int_{-1}^{1} d\widehat{\sigma}(x), \qquad l = 1, ..., n \text{ and } l \text{ is such that } x_{l,n} \neq 0,$$

with $v_{1,l,n}$ being respectively the first components of the normalized eigenvectors of $\mathbf{J}_n^o(I_{q_{n,2}})$, $\mathbf{J}_n^e(I_{q_{n,2}})$ relative to the eigenvalues $x_{l,n}$.

Proof. It remains to determine the sign of $\omega_{2,n}$ in (5.15). This can be done by deriving an explicit formula for $\omega_{2,n}$, from which its sign can be read off. To do so, we first define the polynomial

$$p_{2n}(x) = \frac{q^2_{n,2}(x)}{2(q'_{n,2}(0))^2 + 2q_{n,2}(0)q''_{n,2}(0)}.$$

Therefore since $p_{2n} \in \mathcal{P}_{2n+1}$ and (5.15) is exact on $\mathcal{P}_{2n+1}$, we also have the following representation

$$\omega_{2,n} = \int_{-1}^{1} p_{2n}(x)d\sigma(x).$$

Hence, the desired result now follows immediately, by observing that

$$2(q'_{n,2}(0))^2 + 2q_{n,2}(0)q''_{n,2}(0) = 2(q'_{n,2}(0))^2 > 0$$

if n is odd, and

$$2(q'_{n,2}(0))^2 + 2q_{n,2}(0)q''_{n,2}(0) = -q^2_{n,2}(0)\sum_{l=1}^{n} \frac{1}{x^2_{l,n}} < 0$$

if n is even. □

The following is the basis steps of the numerical algorithm for computing the Turán quadrature formula (5.1).

(1)	Generate the matrix $\mathbf{J}_n(d\widehat{\sigma})$.
(2)	**If (**n is odd) **then**
(3)	Compute d_n^o by (5.6) and use it to generate the matrix $\mathbf{J}_n^o(I_{q_{n,2}})$.
(4)	Compute the eigenvalues and first element of the eigenvectors of $\mathbf{J}_n^o(d\widehat{\sigma})$.
(5)	Compute the coefficients $\omega_{2,n}$ and $\omega_n(0)$ via a direct implementation of (5.16).
(6)	**else**
(7)	Compute d_n^e by (5.12) and use it to generate the matrix $\mathbf{J}_n^e(I_{q_{n,2}})$.
(8)	Compute the eigenvalues and first element the eigenvectors of $\mathbf{J}_n^e(I_{q_{n,2}})$.
(9)	Compute the coefficient $\omega_{2,n}$ via a direct implementation of (5.17).

Of course, in steps (5) and (7), we use the standard software for classical Gaussian quadrature formulae. For a general view, see the recent software package of Gautschi [19].

For a numerical point of view, two types of difficulties still remain about the coefficients necessary for efficient implementation:

a) The step (1) requires the coefficients $\widehat{\beta}_k, k = 1, ..., n-1$. How to evaluate them?

b) The principal difficulty with the steps (3), (5), (7) and (9) lies in the integrals needed to evaluate the exact root d_n of (5.12) and the weights $\omega_{2,n}$, $\omega_n(0)$. How to efficiently compute them?

We give a satisfactory answer to these questions as we will explain below. In the classical cases, the coefficients $\widehat{\beta}_k$ (of the original Jacobi matrix $\mathbf{J}_n(d\widehat{\sigma})$) are known explicitly [42] or are easily derivable from standard results; in others, the software of Gautschi [19] computes them as a preliminary step. For the integrals which are contained in the coefficients d_n^o, d_n^e, $\omega_{2,n}$ and $\omega_n(0)$ they can, in general, not be computed explicitly, to circumvent this difficulty, one judiciously has to use an appropriate quadrature formula to approximate them. We remark that, since the integrands are polynomials, we can evaluate these parameters efficiently to essentially any desired degree of accuracy. However, we shall indicate for the Legendre case how such integrals may be determined explicitly. This issue will be investigated by means of some properties of Jacobi polynomials in the next subsection.

5.2.1. *Legendre case ($d\sigma(x) = dx$).* In the present case the task is easy, since we shall show that the required coefficients of $\mathbf{J}_n(I_{q_{n,2}})$ may be obtained explicitly from $\mathbf{J}_n(d\widehat{\sigma})$. Theorem 10 below describes the calculation of the matrix $\mathbf{J}_n(I_{q_{n,2}})$. For convenience, we first present five technical Lemmas which give us some basic results that will be needed to obtain explicit expressions for (5.1). The proofs of these Lemmas are deferred to the Appendix A in the end of this paper.

To commence our nalysis, we note, first for all, since $d\sigma(x) = dx$, that $d\widehat{\sigma}(x) = x^4 dx$. Thus, in the present case, the orthogonal polynomials $\{\widehat{\pi}_n\}$ can be explicitly written down. In fact, they can be reduced to Jacobi orthogonal polynomials $\left\{p_n^{(\alpha,\beta)}\right\}$ with parameters α, β, (see Szegö [42, formula 4.1.6]):

$$\widehat{\pi}_n(x) = \begin{cases} \left\{ \begin{pmatrix} 2\nu+3/2 \\ \nu \end{pmatrix}^{-1} \right\} p_\nu^{(0,3/2)}(2x^2-1), & \text{if } n = 2\nu, \\ \left\{ \begin{pmatrix} 2\nu+5/2 \\ \nu \end{pmatrix}^{-1} \right\} x p_\nu^{(0,5/2)}(2x^2-1), & \text{if } n = 2\nu+1. \end{cases} \tag{5.18}$$

For convenience, we let

$$c_1(\nu) = \left\{ \begin{pmatrix} 2\nu+3/2 \\ \nu \end{pmatrix}^{-1} \right\} \text{ and } c_2(\nu) = \left\{ \begin{pmatrix} 2\nu+5/2 \\ \nu \end{pmatrix}^{-1} \right\}. \tag{5.19}$$

This representation is the key tool for the analysis we undertake. As an useful consequence, this additional information allows us, for example, to derive explicit formulae for the elements of the matrix $\mathbf{J}_n(d\widehat{\sigma})$. We have, in fact,

Lemma 2. *The elements $\widehat{\beta}_k, k = 1, ..., n-1$, on the side diagonals of the Jacobi matrix $\mathbf{J}_n(d\widehat{\sigma})$ are given by*

$$\widehat{\beta}_k = \begin{cases} 4\nu^2/((4\nu+3)(4\nu+5)), & \text{if } k = 2\nu, \\ (2\nu+5)^2/((4\nu+5)(4\nu+7)), & \text{if } n = 2\nu+1. \end{cases} \tag{5.20}$$

The next Lemma shows that, when n is odd the nodes polynomial $q_{n,2}^o$ (5.5) is simply related to the (monic) Jacobi polynomials.

Lemma 3. *Let n be odd $(n = 2\nu+1)$. Suppose c_2 and $\widehat{\beta}_{2\nu}$ are given respectively as in (5.19) and (5.20). Then the nodes polynomial $q_{n,2}$ of (5.1) has the following representation*

$$q_{n,2}^o(x) = c_2(\nu) x p_\nu^{(0,5/2)}(2x^2-1) + \widehat{\beta}_{2\nu} c_2(\nu-1) x p_{\nu-1}^{(0,5/2)}(2x^2-1). \tag{5.21}$$

While expressions for the weights $\omega_{2,n}$ and $\omega_n(0)$ in (5.1) have already been given in (5.16) and (5.17), for general measure, alternative, more explicit, formulae can be obtained for Legendre measure. Those for the odd case, given in the next Lemma, are particularly simple.

Lemma 4. *Let n be odd $(n = 2\nu+1)$. Then for the Legendre measure the quadrature weights $\omega_{2,n}$ and the one $\omega_n(0)$ relative the term $f(0)$ in (5.1) are given by*

$$\begin{aligned} \omega_{2,n} &= \frac{3\pi}{16} \left\{ \frac{\Gamma(\nu+1)}{\Gamma(\nu+5/2)} \right\}^2, \\ \omega_n(0) &= \frac{6(5+10\nu+4\nu^2)}{5} \omega_{2,n}. \end{aligned} \tag{5.22}$$

The positivity of the weight $\omega_{2,n}$ is evident from (5.22). For the case n even, to simplify the computation of the elements of $\mathbf{J}_n(I_{q_{n,2}})$, the next result identifies the nodes polynomial $q_{n,2}^e$.

Lemma 5. *Let n be even $(n = 2\nu)$. Define*

$$d_n = \frac{\nu(3+2\nu)(5-3\sqrt{5}+(12-4\sqrt{5})\nu+8\nu^2)}{(-1+2\nu+4\nu^2)(3+16\nu+16\nu^2)}.$$

Suppose c_1 are given as in (5.19). Then the nodes polynomial $q_{n,2}$ of (5.1) has the following representation

$$q^e_{n,2}(x) = c_1(\nu)p^{(0,3/2)}_{\nu}(2x^2-1) + d_n c_1(\nu-1)p^{(0,3/2)}_{\nu-1}(2x^2-1). \tag{5.23}$$

For the even case, the weight $\omega_{2,n}$ corresponding the term $f''(0)$ in (5.1) could also be expressed in term of n alone, as shown in he following Lemma.

Lemma 6. *Let n be even $(n = 2\nu)$. Then for the Legendre measure the quadrature weight $\omega_{2,n}$ in (5.1) is given by*

$$\omega_{2,n} = -\frac{\sqrt{5}\pi}{(1+\sqrt{5})^2}\frac{3(3+2\nu)}{8\nu}\left\{\frac{\Gamma(\nu+1)}{\Gamma(\nu+5/2)}\right\}^2$$

We are now in condition to give the principal numerical tool for obtaining the quadrature formula (5.1). This result is an immediate consequence of Lemmas 2, 3, 4, 5, 6 and Theorem 9.

Theorem 10. *Let $\mathbf{J}_n$ be the $(n)-$th order tridiagonal matrix defined by 0 on the main diagonal and $\sqrt{\beta_k}$ on the side diagonals, with*

$$\widehat{\beta}_k = \begin{cases} 4\nu^2/((4\nu+3)(4\nu+5)), & \text{if } k = 2\nu, \\ (2\nu+3)^2/((4\nu+1)(4\nu+3)), & \text{if } k = 2\nu-1. \end{cases}, \ k = 1, ..., n-2,$$

and

$$\widehat{\beta}_{n-1} = \begin{cases} 0, & \text{if } n \ = 2m+1, \\ \dfrac{(2m+3)^2}{(4m+1)(4m+3)} - d_n, & \text{if } n \ = 2m, \end{cases},$$

where

$$d_n = \frac{m(3+2m)(5-3\sqrt{5}+(12-4\sqrt{5})m+8m^2)}{(-1+2m+4m^2)(3+16m+16m^2)}.$$

Then, there exists a unique quadrature formula of the form,

$$\int_{-1}^{1} f(x)dx = \omega^L_{2,n}f''(0) + \sum_{l=1}^{n}\lambda^L_{l,n}f(x^L_{l,n}), \forall f \in \mathcal{P}_{2n+1}, \tag{5.24}$$

where $x^L_{i,n}$, $i = 1, ..., n$, are the n distinct real eigenvalues of J_n and the respective weights are given by

$$\lambda^L_{l,n} = \frac{2}{5}\frac{v^2_{1,l,n}}{x^4_{l,n}}, \qquad l = 1, ..., n \text{ and } l \text{ is such that } x_{l,n} \neq 0,$$

with $v_{1,l,n}$ being the first components of the normalized eigenvectors of $\mathbf{J}_n$ corresponding to the eigenvalues $x_{l,n}$. The weights $\omega^L_{2,n}$ and $\lambda^L_n(0)$ (the weight relative the term $f(0)$) are given by

$$\omega^L_{2,n} = \begin{cases} \dfrac{3\pi}{16}\left\{\dfrac{\Gamma(m+1)}{\Gamma(m+5/2)}\right\}^2, & \text{if } n = 2m+1, \\ -\dfrac{\sqrt{5}\pi}{(1+\sqrt{5})^2}\dfrac{3(3+2m)}{8m}\left\{\dfrac{\Gamma(m+1)}{\Gamma(m+5/2)}\right\}^2 & \text{if } n \ = 2m, \end{cases}$$

$$\lambda^L_n(0) = \frac{9\pi(5+10m+4m^2)}{40}\left\{\frac{\Gamma(m+1)}{\Gamma(m+5/2)}\right\}^2, \ \text{if } n = 2m+1.$$

This quadrature formula enjoys perfect symmetry; indeed we can prove easily that

$$x^L_{n+1-i,n} + x^L_{i,n} = 0,\ \lambda^L_{n+1-i,n} = \lambda^L_{i,n}, \qquad i = 1, ..., n.$$

The construction of quadrature formulae of type (5.1), for a more general weight function, in particular, for Jacobi weight function presents no extra difficulties. The latter can be done in precisely the same way that these results were established for the Gauss-Legendre case. We therefore omit this development here.

5.3. **Comparative numerical results.** In this section, we compare the quadrature formula (BL_n)

$$\int_{-1}^{1} f(x)dx = \omega^L_{2,n} f''(0) + \sum_{l=1}^{n} \lambda^L_{l,n} f(x^L_{l,n}), \forall f \in \mathcal{P}_{2n+1}, \tag{5.25}$$

developed in Theorem 10 and the classical Gauss-Legendre formula (GL_{n+1})

$$\int_{-1}^{1} f(x)dx = \sum_{l=1}^{n+1} \lambda_{l,n} f(x_{l,n}), \forall f \in \mathcal{P}_{2n+1}. \tag{5.26}$$

Note that the quadrature formulae (5.25) and (5.26) use the same number of evaluations of integrand and have the same MPDE. For illustration, we experiment with numerical examples which involve three different kinds of integration problems. They differ mostly in the specific properties of the integrand functions. The first one is a parametrized family of functions which have a peak at a point near the point 0 with a severity that is controlled by a parameter. In the second example, in another context, we will choose a whole family of integrands which have a logarithmic endpoint singularity, and then we conclude with a final example in which the oscillations increase. The three examples are as follows:

$$\begin{aligned} I_1(v) &= \textstyle\int_{-1}^{1} \frac{1}{3\,v^2+(-v+x)^2} dx, \quad (v > 0) \\ &= \frac{\arctan(\frac{1-v}{\sqrt{3}\,v})}{\sqrt{3}\,v} - \frac{\arctan(\frac{-1-v}{\sqrt{3}\,v})}{\sqrt{3}\,v}, \\ I_2(v) &= \textstyle{}_{-1}\frac{\log(1+x)}{\left(1+\left(-\frac{1}{4}+v\right)x\right)^2} dx, \qquad \omega \in (0,1) > 0, \\ &= \frac{16\,\log(2)}{3-8\,v-16\,v^2} - \frac{16\,\log(2)}{5-24\,v+16\,v^2} \\ &\quad - \frac{16\,\log(5-4\,v)}{5-24\,v+16\,v^2} + \frac{16\,\log(3+4\,v)}{5-24\,v+16\,v^2} \end{aligned}$$

$$I_3(m) = \int_0^{2\pi} x\cos(50x)\sin(mx)dx = \frac{2m\pi}{2500-m^2},\ (m \neq 50),$$

All the computations described in this paper were carried out on a personal IBM computer in a double precision.

Example 1. *The integrands here are the functions* $f_v(x) = \frac{1}{3\,v^2+(-v+x)^2}$, $v > 0$, *which satisfy* $f_v'(0) = \dfrac{1}{8v^3}$ *and* $f_v''(x) = \dfrac{6\,x\,(-2\,v+x)}{\left(3\,v^2+(-v+x)^2\right)^3}$. *It should be noted that the graph of* f_v *has the inflection point* $\left(0, \dfrac{1}{4v^2}\right)$, *(this graph is concave upward over* $(-1, 0)$, *and concave downward over* $(0, 2v)$*). It can also be seen that, when* v *is sufficiently small, this function exhibits a peak near* $x = 0$, *with increasing steepness*

m	n	*err.* GL_{n+1}	*err.* BL_n
	130	*8.9(-9)*	*-1.0(-8)*
	180	*2.2(-12)*	*1.2(-12)*
1	*240*	*-6.1(-14)*	*6.3(-14)*
	300	*-1.0(-14)*	*6.3(-14)*
	130	*2.0(-3)*	*-1.7(-3)*
	180	*-2.7(-5)*	*2.1(-5)*
2	*240*	*1.3(-7)*	*-1.1(-7)*
	300	*-7.1(-10)*	*5.5(-10)*
	130	*-1.7*	*1.0*
	180	*2.0(-2)*	*2.3(-2)*
4	*240*	*1.2(-2)*	*-8.3(-3)*
	300	*1.1(-4)*	*-2.1(-4)*

TABLE 1. Numerical results of $I_1(v)$ and comparison with Gauss quadrature.

as v tends to 0. One would expect that the difficulty of the integrand f depends heavily on the selection of the parameter v, which is equal to the location of the peak. For this example, since $f_v''(0) = 0$, it seems reasonable, therefore, to compare the quadrature formula (BL_n) with the ordinary Gaussian quadrature (GL_n) having the same number n of interior nodes. We did our computation for $v = v_m = 1/10^{2m}$, $m = 1, ..., 6$ and $n = 130, 180, 240, 300$, but in Table 1 shown only select results. Also shown in the last two columns are the respective integration errors. It can be seen Table 1 that the new quadrature formula produces more accurate result than those furnished by the ordinary Gauss quadrature formula, especially for v close 0. With decreasing v, ordinary Gauss-Legendre is seen to converge rather slowly. This weakness is accentuated when the peak is moderately high. However, for small m, both quadratures produce results which are comparable in accuracy, the formula (GL) being somewhat more accurate than (BL) for small values n, the latter for larger values of n.

Example 2. *The integrand here is a function having a logarithmic endpoint singularity of the type $f_v(x) = \dfrac{\log(1+x)}{\left(1+\left(-\frac{1}{4}+v\right)\,x\right)^2}$. Thus, $f_v(0) = 0$, $f_v'(0) = 0$ and $f_v''(0) = -4v$. We therefore compare the results of (5.25), at the same cost, by the ordinary $(n+1)-$Gauss quadrature formula (5.26). The results are shown in Table 2 for selected values of $v = 1/(100j)$ and n. It can be seen that both formulae converge slowly, with (BL) having a slight edge on (GL), particulary when v is very close to 0.*

Example 3. *As suggested in the introducation, we can compare the results of the composite quadrature formula,*

$$CBL_n(f) = \sum_{i=1}^{r}\sum_{j=1}^{s_i}\left(w_{j,s_i}^{BL} f''(0) + w_{j,i,s}^{BL} f(x_{j,i,s_i}^{BL})\right),$$

j	n	*err.* GL_{n+1}	*err.* BL_n
	50	*3.0(-4)*	*3.2(-4)*
1	*150*	*3.6(-5)*	*3.6(-5)*
	250	*1.2(-5)*	*1.3(-5)*
	50	*3.3(-4)*	*3.4(-4)*
2	*150*	*3.8(-5)*	*3.8(-5)*
	250	*1.3(-6)*	*1.3(-6)*
	50	*3.6(–4)*	*3.7(-4)*
3	*150*	*4.1(-5)*	*3.2(-6)*
	250	*1.5(-5)*	*1.2(-7)*

TABLE 2. Numerical results of $I_2(w)$ and comparison with Gauss quadrature.

n	*err.* GL_{n+1}	*err.* BL_n
200	*1.02(-7)*	*9.9(-8)*
250	*2.18(-12)*	*3.581(-13)*

TABLE 3. Numerical results of $I_1(v)$ and comparison with Gauss quadrature.

n	*err.* GL_{n+1}	*err.* BL_n
200	*1.17(-6)*	*8.93(-7)*

TABLE 4. Numerical results of $I_1(v)$ and comparison with Gauss quadrature.

with the standard composite quadrature formula

$$CGL_{n+1}(f) = \sum_{i=1}^{r+1} \sum_{j=1}^{s_i+1} w_{j,i,s}^{GL} f(x_{j,i,s}^{GL})$$

relative to the same subdivision $[-1,1] = \bigcup_{i=1}^{r} [\xi_{i-1}, \xi_i]$. *We illustrate these formulae by computing again the integrals* $I_i(v), i = 1, 2$. *We will refer to the number used by each quadrature as* $n = \sum_{i=1}^{r} s_i$. *For* $I_1(v)$, *we did our computation for* $v = 1/10^4, r = 3$, $s_1 = 50$, $s_2 = 100, 150$, $s_3 = 50$ *and therefore* $n = 200, 250$. *It seems reasonable to employ the quadrature formulae* CBL_n *and* CGL_{n+1} *choosing the nodes* $\zeta_i, i = 1, ..., r$, *near* 0. *Table 3 shows the respective integration errors for* $\zeta_0 = -1, \zeta_1 = -0.2, \zeta_2 = 0.2$ *and* $\zeta_2 = 1$.

Table 3 shows the error of the s-point approximations $I_2(v)$, *for* $v = 1/300, r = 3$, $s_1 = 90$, $s_2 = 70$, $s_3 = 40$ *and therefore* $n = 200$.

These examples demonstrate the superiority of CBL_n *over* CGL_{n+1}. *However, both quadrature formulae are even better than* BL_n *and* GL_{n+1}. *For example, if* $I_2(1/10^4)$ *is approximated by* BL_n *or* GL_{n+1} *with* $n = 300$, *they produce only* 4 *correct decimal digits, which is several orders of magnitude larger that* BL_n *and* GL_{n+1}.

6. Applications to polynomial inequalities

In this last section we apply the results of the previous sections to derive and calculate, in a simple manner, new explicit sharp weighted L^2 inequalities for coefficients of polynomials. They are all obtained as direct applications of the quadrature formula

$$\int_{-1}^{1} f(x)d\sigma(x) = \sum_{i=0}^{m} \omega_{i,n} f^{(2i)}(0) + \sum_{l=1}^{n-m} \lambda_{l,m,n} f(x_{l,m,n}), \quad \forall f \in \mathcal{P}_{2n+1}. \tag{6.1}$$

We point out that the nodes of such quadrature formula are the zeros of the orthogonal polynomials associated with respect to the measure $d\widehat{\sigma} = x^{2m+2}d\sigma(x)$. This result can be obtained as an immediate consequence of Theorem 5.

Problem 3. *Let $p(x) = a_0+a_1x+...+a_nx^n$ be a polynomial of degree n. We assume that $\|p\|_{m,2} = \left(\int_{-1}^{1} x^{2m}\left\{\frac{d^m}{dx^m}p(x)\right\}^2 d\sigma(x)\right)^{1/2}$ is given as a measure of the "size" of p. We now ask the question, how large the coefficient $|a_m|$ can be? Below, we answer this question. We will calculate explicitly the constants that appear in our estimations. These are derived using essentially the quadrature formula (6.1).*

Before stating the particular Legendre case, we prove a more general theorem which leads to the best possible estimation for $|a_m|$ in terms of $\|p\|_{m,2}$.

Theorem 11. *Let $p(x) = \sum_{i=0}^{n} a_i x^i$ be a polynomial of degree n. Then for any fixed $k, 0 \leq m \leq n$, such that $n-m$ is even, we have*

$$|a_m| \leq \frac{1}{m!} \frac{\left\|\widehat{L}_{n-m}\right\|_{m,2}}{\left|\int_{-1}^{1} x^{2m}\widehat{L}_{n-m}(x)d\sigma(x)\right|} \|p\|_{m,2}, \tag{6.2}$$

where $\widehat{L}_{n-m}(x) = \prod_{l=1}^{n-m}(x - x_{l,m,n})$, with $x_{l,m,n}, l = 1, ..., n$, being the nodes of the quadrature formula (6.1). The equality hods if and only if $\frac{d^m}{dx^m}p(x) = const\ \widehat{L}_{n-m}(x)$.

Remark 3. *An equivalent expression of (6.2) is*

$$\frac{\left\|\widehat{L}_{n-m}\right\|_{m,2}}{m!\left|\int_{-1}^{1} x^{2m}\widehat{L}_{n-m}(x)d\sigma(x)\right|} = \max\left\{\left|\frac{d^m}{dx^m}p(0)\right| : p \in \mathcal{P}_n, \|p\|_{m,2} = 1\right\}.$$

More explicit estimates than those in Theorem 11 can be obtained when the orthogonal polynomials associated with respect to the measure $d\widehat{\sigma} = x^{2m+2}d\sigma(x)$ can be explicitly written down. The content of the following result is to give an exact expression of the above constant, for simplicity, we only consider the Legendre case. Thus, if $d\sigma(x) = dx$ then the above Theorem becomes the following result, which will be proven in the Appendix B.

Corollary 2. *Let* $p(x) = \sum_{i=0}^{n} a_i x^i$ *be a polynomial of degree* n. *Then for any fixed* $k, 0 \le k \le n$, *such that* $n - m = 2\nu$, *we have*

$$|a_m| \le \frac{1}{(m!)^2} \frac{\Gamma(\nu + m + 3/2)}{\sqrt{m + 1/2}\ \Gamma(m + 1/2)} \|p\|_{m,2}\,, \tag{6.3}$$

The equality hods if and only if $\dfrac{d^m}{dx^m} p(x) = const\ p_\nu^{(0,m+1/2)}(2x^2 - 1)$.

Let us recall that a number of weighted L^2 inequalities for coefficients of polynomials have been suggested in [31] and in [25]. A full history and a detailed account of the current state of this theory and its applications appears in the recent monograph [35].

7. Concluding remarks

Our purpose has been to shown how to modify the Jacobi matrix in such a way that the nodes and weights of a new class of quadratures can then be evaluated directly by standard software, the complexity of the various calculations associated with computing such quadratures is significantly reduced. Particularly encouraging is that this matrix is easy to construct employing only elementary transformations of the recurrence coefficients which characterize the original measure. This approach extends the well-known method for ordinary Gaussian quadrature formulae. The formulae derived in this way have been implemented and some numerical tests performed with the latter show a improvement over the classical Gaussian quadrature formulae. We have also exploited the connection between quadrature formulae and polynomials inequalities to apply the new class of quadratures for the latter to derive new explicit sharp inequalities for polynomials.

For sake of simplicity only even measure has been considered, but the method developed here can be applied to arbitrary measure with slight technical modifications. It can also be extended to the construction of quadratures with more general *interior* conditions, such as the following class

$$Q(f) = \sum_{j=1}^{k} \omega_{j,n} C_j(f) + \sum_{i=1}^{n} \lambda_{i,n} f(x_{i,n}),$$

where C_j, $j = 1, ..., k$, are given more general linear functionals of the form

$$C_l(f) := \sum_{m=0}^{q_l - 1} a_{lm} f^{(m)}(0), \qquad l = 1, ..., k.$$

The analysis leading to these new quadrature formulae takes place most naturally on more general domains, for example, on semi-finite or infinite with adapted interior conditions. This problem requires different approaches, to which we hope to return.

Finally, the possibly most important merit of this class of quadratures is that they can be useful in the numerical solution by spectral approximation of partial differential equations, involving *interior* conditions, as has been successfully applied in the author's joint work [17], where efficient methods based on a new family of quadrature formulae, with complex *boundary* conditions terms, were proposed. In that paper, the appending of additional terms in the quadratures was motived by the well known lumped mass spectral approximation problem.

We have also apply these quadratures to derive and calculate new explicit sharp weighted L^2 inequalities for coefficients of polynomials.

Appendix A: Proofs of Lemmas. In this Appendix, we provide the proofs of Lemmas 2, 3, 4, 5 and 6. The proofs involve substantial use of the properties of Jacobi polynomials, these properties are first summarized. They are included here to facilitate verification by the interested reader. All the results are standard and can be found in books such that [42, 3].

We start with some basic notations and known results. For any real number a and integer n, the Prochhammer's symbol $(a)_n$ is defined by

$$(a)_n = a(a+1)...(a+n-1) = \frac{\Gamma(a+n)}{\Gamma(a)},$$

where Γ is the Gamma function. With the weight function

$$d\sigma^{\alpha,\beta}(x) = (1-x)^{\alpha}(1+x)^{\beta} \qquad \text{on } (-1,1), \qquad (\alpha,\beta > -1),$$

we associate the scalar product

$$(\phi,\psi)_{\alpha,\beta} := \int_{-1}^{1} \phi(x)\psi(x)d\sigma^{\alpha,\beta}(x), \tag{7.1}$$

which is defined on the space of all functions whose square is integrable with respect to the measure $d\sigma^{\alpha,\beta}$. There exists a sequence of polynomials called Jacobi polynomials of parameters α and β, which are orthogonal w. r. t. the scalar product $(.,.)_{\alpha,\beta}$. As in [42], we shall use for these polynomials, the (unusual) notation $p_n^{(\alpha,\beta)}$ and normalization

$$p_n^{(\alpha,\beta)}(1) = \frac{(\alpha+1)_n}{n!}. \tag{7.2}$$

Their orthogonality relation is [42, Formula 4.3.3]

$$(p_n^{(\alpha,\beta)}, p_m^{(\alpha,\beta)})_{\alpha,\beta} = h_n\delta_{nm}, \tag{7.3}$$

where

$$h_n = \frac{2^{\alpha+\beta+1}\Gamma(n+\alpha+1)\Gamma(n+\beta+1)}{(2n+\alpha+\beta+1)\Gamma(n+1)\Gamma(n+\alpha+\beta+1)},$$

and δ_{nm} is the Kronecker delta.

They have the explicit representation (see [42, Formula 4.3.2])

$$p_n^{(\alpha,\beta)}(x) = \sum_{\nu=0}^{n} \begin{pmatrix} n+\alpha \\ n-\nu \end{pmatrix} \begin{pmatrix} n+\beta \\ \nu \end{pmatrix} \left(\frac{x-1}{2}\right)^{\nu} \left(\frac{x+1}{2}\right)^{n-\nu}.$$

Note that

$$p_n^{(\alpha,\beta)}(x) = \frac{(n+\alpha+\beta+1)_n}{2^n n!}x^n + \text{lower order terms}. \tag{7.4}$$

This gives us that the associated (monic) orthogonal polynomials are given by

$$\widehat{p}_n^{(\alpha,\beta)}(x) = \frac{2^n n!}{(n+\alpha+\beta+1)_n} p_n^{(\alpha,\beta)}(x).$$

Further we have

$$p_n^{(\alpha,\beta)}(-x) = (-1)^n p_n^{(\beta,\alpha)}(x). \tag{7.5}$$

As it is well known for Jacobi polynomials the derivatives of arbitrary order are also Jacobi polynomials. This is easily derived from general properties of the hypergeometric functions $F(a;b;c;z)$ [42]. For instance, we have

$$p_n^{(\alpha,\beta)}(1-2z) = \frac{(\alpha+1)_n}{n!} F(-n, n+\alpha+\beta+1; \alpha+1; z).$$

Since

$$\frac{d}{dz}F(a;b;c;z) = \frac{ab}{c}F(a+1;b+1;c+1;z),$$

it follows the simple differentiation formula

$$\frac{d}{dx}p_n^{(\alpha,\beta)}(x) = \frac{1}{2}(n+\alpha+\beta+1)p_{n-1}^{(\alpha+1,\beta+1)}(x),$$

or more generally for $k \leq n$

$$\frac{d^k}{dx^k}p_n^{(\alpha,\beta)}(x) = \frac{1}{2^k}\frac{\Gamma(n+k+\alpha+\beta+1)}{\Gamma(n+\alpha+\beta+1)}p_{n-k}^{(\alpha+k,\beta+k)}(x). \tag{7.6}$$

Finally, we will need to use heavily the following formulae, see [42, Formula 4.5.4, p. 72], [3, Formula 4, p. 263] and [3, Formula 5, p. 263].

$$(1+x)\ p_n^{(\alpha,\beta+1)}(x) = \frac{2}{2n+\alpha+\beta+2}\left((n+\beta+1)p_n^{(\alpha,\beta)}(x) + (n+1)p_{n+1}^{(\alpha,\beta)}(x)\right), \tag{7.7}$$

$$\int_{-1}^{1} p_n^{(\alpha,\beta)}(x)d\sigma^{\rho,\beta}(x) = \frac{2^{\beta+\rho+1}\Gamma(\rho+1)\Gamma(\beta+n+1)\Gamma(\alpha-\rho+n)}{n!\Gamma(\alpha-\rho)\Gamma(\beta+\rho+n+2)},\ \rho<\alpha, \tag{7.8}$$

and

$$\int_{-1}^{1}\left\{p_n^{(\alpha,\beta)}(x)\right\}^2 d\sigma^{\alpha-1,\beta}(x) = \frac{2^{\alpha+\beta}\Gamma(\alpha+n+1)\Gamma(\beta+n+1)}{n!\alpha\Gamma(\alpha+\beta+n+1)}. \tag{7.9}$$

Now we prove the technical results needed in the proof of Theorem 10. We begin with the proof of Lemma 2.

Proof. Let first k be even ($k=2\nu$). Then, by (4.3) and (5.18) we have

$$\begin{aligned}
\widehat{\beta}_k &= \frac{c_1^2(\nu)}{c_2^2(\nu)}\frac{\int_{-1}^{1} x^4\left\{p_\nu^{(0,3/2)}(2x^2-1)\right\}^2 dx}{\int_{-1}^{1} x^6\left\{p_{\nu-1}^{(0,5/2)}(2x^2-1)\right\}^2 dx} \\
&= \frac{c_1^2(\nu)}{c_2^2(\nu)}\frac{\int_{0}^{1} x^4\left\{p_\nu^{(0,3/2)}(2x^2-1)\right\}^2 dx}{\int_{0}^{1} x^6\left\{p_{\nu-1}^{(0,5/2)}(2x^2-1)\right\}^2 dx} \\
&= \frac{2c_1^2(\nu)}{c_2^2(\nu)}\frac{\int_{-1}^{1}(1+u)^{3/2}\left\{p_\nu^{(0,3/2)}(u)\right\}^2 du}{\int_{-1}^{1}(1+u)^{5/2}\left\{p_{\nu-1}^{(0,5/2)}(u)\right\}^2 du},
\end{aligned}$$

where in the last equality, we have used the change of variable $u = 2x^2-1$. Since, from (7.3) we have

$$\int_{-1}^{1}(1+u)^{3/2}\left\{p_\nu^{(0,3/2)}(u)\right\}^2 du = \frac{2^{5/2}}{(2\nu+5/2)},$$

and

$$\int_{-1}^{1}(1+u)^{5/2}\left\{p_\nu^{(0,3/2)}(u)\right\}^2 du = \frac{2^{7/2}}{(2\nu+3/2)},$$

one immediately obtains

$$\widehat{\beta}_k = \frac{(2\nu+3/2)c_1^2(\nu)}{(2\nu+5/2)c_2^2(\nu)}. \tag{7.10}$$

On the other hand, we have

$$\frac{c_1(\nu)}{c_2(\nu)} = \frac{\nu}{(2\nu+3/2)}.$$

Inserted in (7.10), this yields the first relation in (5.20). The case k even is proved similarly. □

We add a remark concerning more general nonnegative even measure $d\sigma$: In the present case, one may verify directly that the polynomial $x\widehat{\pi}_{n-1}$ $(= q^o_{n,2})$ has the matrix representation $x\widehat{\pi}_{n-1}(x) = \det(xI_n - \mathbf{J}_n)$, where

$$\mathbf{J}_n = \begin{bmatrix} 0 & \sqrt{\widehat{\beta}_1} & 0 & \dots & 0 \\ \sqrt{\widehat{\beta}_1} & 0 & \ddots & \ddots & \vdots \\ 0 & \ddots & \ddots & \sqrt{\widehat{\beta}_{n-1}} & 0 \\ \vdots & \ddots & \sqrt{\widehat{\beta}_{n-1}} & 0 & 0 \\ 0 & \cdots & 0 & 0 & 0 \end{bmatrix}_{n\times n}.$$

A trivial consequence, which, by applying Theorem 7, could also be used to show (5.21), is that $d_n = \widehat{\beta}_{n-1}$.

Remark 4. *It is remarkable that, taking into account that we also have $q^o_{n,2}(x, d\sigma) = x\widehat{\pi}_{n-1}(x, d\sigma)$, $\forall$ $d\sigma$, the particular and convenient above structure of the matrix $\mathbf{J}_n$ remains valid more generally, for any nonnegative even measure $d\sigma$.*

We now proceed to the proof of Lemma 3.

Proof. We need only compute $d^o_n = \int_{-1}^1 x\widehat{\pi}_{n-1}(x)\widehat{\pi}_{n-2}(x)d\widehat{\sigma}(x)/\int_{-1}^1 \widehat{\pi}^2_{n-2}(x)d\widehat{\sigma}(x)$, since, from (5.5) and (5.18), we have

$$q^o_{n,2}(x) = c_2(\nu)xp_\nu^{(0,5/2)}(2x^2-1) + d^o_n c_2(\nu-1)xp_{\nu-1}^{(0,5/2)}(2x^2-1).$$

For this, note first that

$$\begin{aligned} d^o_n &= \frac{c_1(\nu)}{c_2(\nu)} \frac{\int_{-1}^1 x^4 \left\{p_\nu^{(0,3/2)}(2x^2-1)p_{\nu-1}^{(0,5/2)}(2x^2-1)\right\} dx}{\int_{-1}^1 x^6 \left\{p_{\nu-1}^{(0,5/2)}(2x^2-1)\right\}^2 dx} \\ &= \frac{c_1(\nu)}{c_2(\nu)} \frac{\int_{0}^1 x^4 \left\{p_\nu^{(0,3/2)}(2x^2-1)p_{\nu-1}^{(0,5/2)}(2x^2-1)\right\} dx}{\int_{0}^1 x^6 \left\{p_{\nu-1}^{(0,5/2)}(2x^2-1)\right\} dx} \\ &= \frac{c_1(\nu)}{c_2(\nu)} \frac{\int_{-1}^1 (1+u)^{5/2} \left\{p_\nu^{(0,3/2)}(u)p_{\nu-1}^{(0,5/2)}(u)\right\} du}{\int_{-1}^1 (1+u)^{5/2} \left\{p_{\nu-1}^{(0,5/2)}(u)\right\}^2 du}. \end{aligned} \tag{7.11}$$

From(7.3) we have

$$\int_{-1}^1 (1+u)^{5/2} \left\{p_\nu^{(0,5/2)}(u)\right\}^2 du = \frac{2^{7/2}}{(2\nu+3/2)}. \tag{7.12}$$

The formula (7.7), with $\alpha = 0, \beta = 3/2$ and $n = \nu - 1$, yields

$$(1+u)p_{\nu-1}^{(0,5/2)}(u) = \frac{2}{2\nu+3/2}\left((\nu+3/2)p_{\nu-1}^{(0,3/2)}(u) + \nu p_{\nu}^{(0,3/2)}(x)\right).$$

Therefore, one finds that

$$\int_{-1}^{1}(1+u)^{5/2}\left\{p_{\nu}^{(0,3/2)}(u)p_{\nu-1}^{(0,5/2)}(u)\right\}du = e\int_{-1}^{1}(1+u)^{3/2}\left\{p_{\nu}^{(0,3/2)}(u)\right\}^2 du,$$

where $e = \dfrac{2\nu}{2\nu+3/2}$. Using this, and (7.3) with $\alpha = 0, \beta = 3/2$ and $n = \nu$, gives

$$\int_{-1}^{1}(1+u)^{5/2}\left\{p_{\nu}^{(0,3/2)}(u)p_{\nu-1}^{(0,5/2)}(u)\right\}du = \frac{\nu}{(2\nu+3/2)(2\nu+5/2)}2^{7/2}.$$

This together with (7.12) gives

$$\frac{\int_{-1}^{1}(1+u)^{5/2}\left\{p_{\nu}^{(0,3/2)}(u)p_{\nu-1}^{(0,5/2)}(u)\right\}du}{\int_{-1}^{1}(1+u)^{5/2}\left\{p_{\nu-1}^{(0,5/2)}(u)\right\}^2 du} = \frac{\nu}{(2\nu+5/2)}.$$

Inserted in (7.11), and also using the fact that $\dfrac{c_1(\nu)}{c_2(\nu)} = \dfrac{\nu}{(2\nu+3/2)}$, this proves the result. □

We pass now the proof of Lemma 4.

Proof. It follows easily from (5.16) and (5.18) that

$$\omega_{2,n} = \frac{\sqrt{2}}{8}\frac{\int_{-1}^{1}\sqrt{1+u}\left\{p_{\nu}^{(0,3/2)}(u)\right\}du}{\left\{p_{\nu}^{(0,3/2)}(-1)\right\}},$$

This together with (7.5) and by a change of variable gives

$$\omega_{2,n} = \frac{\sqrt{2}}{8}\frac{\int_{-1}^{1}\sqrt{1-u}\left\{p_{\nu}^{(3/2,0)}(u)\right\}du}{\left\{p_{\nu}^{(3/2,0)}(1)\right\}}. \tag{7.13}$$

Since, from (7.2), we have

$$p_{\nu}^{(3/2,0)}(1) \quad = \quad \frac{\Gamma(\nu+5/2)}{\Gamma(\nu+1)\Gamma(5/2)}, \tag{7.14}$$

it remains to evaluate the integral in the numerator of (7.13). To do this, once again, we appeal to the identity (7.8), with $\rho = 1/2, \alpha = 3/2$ and $\beta = 0$, to obtain

$$\int_{-1}^{1}\sqrt{1-u}\left\{p_{\nu}^{(3/2,0)}(u)\right\}du = 2^{3/2}\frac{\Gamma(\nu+1)}{\Gamma(\nu+5/2)}\Gamma(3/2). \tag{7.15}$$

Finally, combining (7.15), (7.14), (7.13) and using the fact that $\{\Gamma(3/2)\Gamma(5/2)\} = \frac{3\pi}{8}$ yields (5.22). In order to prove the explicit expression for $\omega_n(0)$ we have to use an argument similar the one above. This will be omitted here. □

Next we prove Lemma 5.

Proof. For simplicity, we refer to $d = d_n$. First, (5.11) gives that d is a zero of

$$\frac{\int_{-1}^{1} x^2 \widehat{\pi}_n(x)dx + d\int_{-1}^{1} x^2 \widehat{\pi}_{n-2}(x)dx}{\int_{-1}^{1} \widehat{\pi}_n(x)dx + d\int_{-1}^{1} \widehat{\pi}_{n-2}(x)dx} = 2\frac{\widehat{\pi}_n(0) + d\widehat{\pi}_{n-2}(0)}{\widehat{\pi}_n''(0) + d\widehat{\pi}_{n-2}''(0)}. \tag{7.16}$$

As suggested in the remark after the equation (5.13), to establish the lemma, we need to find the zero d of (7.16) with smaller magnitude. To solve equation (7.16) we first remark that this is equivalent to

$$\frac{1 + d\int_{-1}^{1} x^2 \widehat{\pi}_{n-2}(x)dx/\int_{-1}^{1} x^2 \widehat{\pi}_n(x)dx}{1 + d\int_{-1}^{1} \widehat{\pi}_{n-2}(x)dx/\int_{-1}^{1} \widehat{\pi}_n(x)dx} = 2f\frac{1 + d\widehat{\pi}_{n-2}(0)/\widehat{\pi}_n(0)}{1 + d\widehat{\pi}_{n-2}''(0)/\widehat{\pi}_n''(0)}, \tag{7.17}$$

where

$$f = \frac{\widehat{\pi}_n(0)}{\widehat{\pi}_n''(0)}\frac{\int_{-1}^{1} \widehat{\pi}_n(x)dx}{\int_{-1}^{1} x^2 \widehat{\pi}_n(x)dx}.$$

We chose this representation because it gives a particularly easy proof of Lemma 5. We now show that

$$N_1 := \frac{\int_{-1}^{1} x^2 \widehat{\pi}_{n-2}(x)dx}{\int_{-1}^{1} x^2 \widehat{\pi}_n(x)dx} = -\frac{(2v + 1/2)(2v + 3/2)}{v^2}, \tag{7.18}$$

This follows by observing, since $n = 2\nu$, that

$$\begin{aligned}
\frac{\int_{-1}^{1} x^2 \widehat{\pi}_{n-2}(x)dx}{\int_{-1}^{1} x^2 \widehat{\pi}_n(x)dx} &= \frac{c_1(\nu-1)}{c_1(\nu)}\frac{\int_{-1}^{1} x^2 \left\{p_{\nu-1}^{(0,3/2)}(2x^2-1)\right\}dx}{\int_{-1}^{1} x^2 \left\{p_{\nu}^{(0,3/2)}(2x^2-1)\right\}dx} \\
&= \frac{c_1(\nu-1)}{c_1(\nu)}\frac{\int_{0}^{1} x^2 \left\{p_{\nu-1}^{(0,3/2)}(2x^2-1)\right\}dx}{\int_{0}^{1} x^2 \left\{p_{\nu}^{(0,3/2)}(2x^2-1)\right\}dx} \\
&= -\frac{c_1(\nu-1)}{c_1(\nu)}\frac{\int_{-1}^{1} \sqrt{1-u}\left\{p_{\nu-1}^{(3/2,0)}(u)\right\}du}{\int_{-1}^{1} \sqrt{1-u}\left\{p_{\nu}^{(3/2,0)}(u)\right\}dx},
\end{aligned} \tag{7.19}$$

The result now follows by using (7.15).

To evaluate $\int_{-1}^{1} \widehat{\pi}_{n-2}(x)dx/\int_{-1}^{1} \widehat{\pi}_n(x)dx$ we proceed as follows:

$$\begin{aligned}
\frac{\int_{-1}^{1} \widehat{\pi}_{n-2}(x)dx}{\int_{-1}^{1} \widehat{\pi}_n(x)dx} &= \frac{c_1(\nu-1)}{c_1(\nu)}\frac{\int_{-1}^{1} \left\{p_{\nu-1}^{(0,3/2)}(2x^2-1)\right\}dx}{\int_{-1}^{1} \left\{p_{\nu}^{(0,3/2)}(2x^2-1)\right\}dx} \\
&= \frac{c_1(\nu-1)}{c_1(\nu)}\frac{\int_{0}^{1} \left\{p_{\nu-1}^{(0,3/2)}(2x^2-1)\right\}dx}{\int_{0}^{1} \left\{p_{\nu}^{(0,3/2)}(2x^2-1)\right\}dx} \\
&= -\frac{c_1(\nu-1)}{c_1(\nu)}\frac{\int_{-1}^{1} \left\{p_{\nu-1}^{(3/2,0)}(u)/\sqrt{1-u}\right\}du}{\int_{-1}^{1} \left\{p_{\nu}^{(3/2,0)}(u)/\sqrt{1-u}\right\}dx},
\end{aligned}$$

which, since

$$\int_{-1}^{1} \left\{p_{\nu}^{(3/2,0)}(u)/\sqrt{1-u}\right\}dx = \sqrt{2}\frac{\Gamma(1/2)\Gamma(\nu+2)}{\Gamma(\nu+23/)},$$

reduces to

$$N_2 := \frac{\int_{-1}^{1} \widehat{\pi}_{n-2}(x)dx}{\int_{-1}^{1} \widehat{\pi}_n(x)dx} = -\frac{(v+1/2)(2v+1/2)(2v+3/2)}{v(v+1)(v+3/2)}. \tag{7.20}$$

Now using (7.2), we get from (7.5)

$$D_1 := \frac{\widehat{\pi}_{n-2}(0)}{\widehat{\pi}_n(0)} = -\frac{(2v+1/2)(2v+3/2)}{(v+3/2)^2}. \tag{7.21}$$

To evaluate $\widehat{\pi}''_{n-2}(0)/\widehat{\pi}''_n(0)$, we obtain from

$$\widehat{\pi}_n(x) = c_1(\nu)p_\nu^{(0,3/2)}(2x^2-1)$$

by differentiation this expression twice, then set $x = 0$ and using (7.6)

$$\widehat{\pi}''_n(0) = (-1)^{\nu-1}2(\nu+5/2)c_1(\nu)\frac{\Gamma(\nu+5/2)}{\Gamma(7/2)\Gamma(\nu)},$$

so that

$$D_2 := \frac{\widehat{\pi}''_{n-2}(0)}{\widehat{\pi}''_n(0)} = \frac{(\nu-1)(2v+1/2)(2v+3/2)}{\nu(v+3/2)(v+3/2)}. \tag{7.22}$$

Combining (7.18), (7.20), (7.21) and (7.22) yields

$$\frac{1-N_1\ d}{1-N_2\ d} = 2f\frac{1-D_1\ d}{1-D_2\ d},$$

where

$$f = -\frac{5(v+1)(v+3/2)}{2v(v+5/2)}.$$

Finally, one finds, after a somewhat laborious but elementary calculation, that

$$d = \frac{\nu(3+2\nu)(5-3\sqrt{5}+(12-4\sqrt{5})\nu+8\nu^2)}{(-1+2\nu+4\nu^2)(3+16\nu+16\nu^2)}.$$

The representation (5.23) now follows from (5.13). □

We now can prove Lemma 6.

Proof. By (5.17), with the use of (5.23), we obtain

$$\omega_{2,n} = \frac{1}{2}\frac{\int_{-1}^{1} x^2\widehat{\pi}_n(x)dx + d_n\int_{-1}^{1} x^2\widehat{\pi}_{n-2}(x)dx}{\widehat{\pi}_n(0)+d_n\widehat{\pi}_{n-2}(0)}.$$

It turn out to be convenient for the computation to write

$$\omega_{2,n} = \frac{1}{2}f\frac{1+d_n\int_{-1}^{1} x^2\widehat{\pi}_{n-2}(x)dx/\int_{-1}^{1} x^2\widehat{\pi}_n(x)dx}{1+d_n\widehat{\pi}_{n-2}(0)/\widehat{\pi}_n(0)},$$

where

$$f = \frac{\int_{-1}^{1} x^2\widehat{\pi}_n(x)dx}{\widehat{\pi}_n(0)}.$$

Since

$$\frac{1+d_n\int_{-1}^{1} x^2\widehat{\pi}_{n-2}(x)dx/\int_{-1}^{1} x^2\widehat{\pi}_n(x)dx}{1+d_n\widehat{\pi}_{n-2}(0)/\widehat{\pi}_n(0)} = \frac{1-N_1d_nd}{1-D_1\ d_n},$$

where N_1 and D_1 are given by (7.18) and (7.21), we then have

$$\omega_{2,n} = \frac{1}{2}f\frac{1-N_1\ d_n}{1-D_1\ d_n}. \tag{7.23}$$

We also have

$$\begin{aligned}\frac{\int_{-1}^{1} x^2 \widehat{\pi}_n(x)dx}{\widehat{\pi}_n(0)} &= \frac{\int_{-1}^{1} x^2 \left\{p_\nu^{(0,3/2)}(2x^2-1)\right\} dx}{\left\{p_\nu^{(0,3/2)}(-1)\right\}} \\ &= 2\frac{\int_{0}^{1} x^2 \left\{p_\nu^{(0,3/2)}(2x^2-1)\right\} dx}{\left\{p_\nu^{(0,3/2)}(-1)\right\}} \\ &= \frac{\sqrt{2}}{4}\frac{\int_{-1}^{1} \sqrt{1-u}\left\{p_\nu^{(3/2,0)}(u)\right\} du}{\left\{p_\nu^{(3/2,0)}(1)\right\}},\end{aligned}$$

then we use (7.15) and (7.2) to obtain

$$f = \frac{3}{2}\left\{\frac{\Gamma(3/2)\Gamma(\nu+1)}{\Gamma(\nu+5/2)}\right\}^2,$$

which, inserted in (7.23) and using the value of d_n found in Lemma (5), yields after a simple computation the desired result. □

Appendix B: Proof of Theorem 11 and Corollary 2.

The technique we use in the proofs is essentially the same as that in [8, 26]. We begin by the proof of *Theorem* **11.**

Proof. This is an immediate consequence of the quadrature formula (6.1) and the Cauchy-Schwarz inequality, combined with a cleaver idea that appears in [8]. By the quadrature formula (6.1) which is exact in $\mathcal{P}_{2n+1}$, we have for any $p \in \mathcal{P}_n$

$$\int_{-1}^{1} x^{2m}\left\{\frac{d^m}{dx^m}p(x)\right\}\widehat{L}_{n-m}(x)d\sigma(x) = (2m)!\left\{\frac{d^m}{dx^m}p(0)\right\}\widehat{L}_{n-m}(0)\omega_{m,n}.$$

Hence, using the fact that

$$\int_{-1}^{1} x^{2m}\widehat{L}_{n-m}(x)d\sigma(x) = (2m)!\widehat{L}_{n-m}(0)\omega_{m,n},$$

we get

$$a_m = \frac{1}{m!}\frac{\int_{-1}^{1} x^{2m}\left\{\frac{d^m}{dx^m}p(x)\right\}\widehat{\pi}_{n-m}(x)d\sigma(x)}{\int_{-1}^{1} x^{2m}\widehat{\pi}_{n-m}(x)d\sigma(x)}.$$

The theorem now follows from the Cauchy-Schwarz inequality. The sharpness of the estimate is guaranteed by the latter. □

We now mention how we can prove *Corollary 2*

Proof. When $d\sigma(x) = dx$, using the same arguments as before, the inequality (6.3) is a simple combination of Theorem **11,** the identities (7.8), (7.9) and the fact that

$$\widehat{\pi}_{n-m}(x) = \left\{\begin{pmatrix} 2\nu+m+1/2 \\ \nu \end{pmatrix}^{-1}\right\} p_\nu^{(0,m+1/2)}(2x^2-1), \qquad (n-m=2\nu).$$

Details are omitted for brevity. □

Acknowledgment. The author would like to acknowledge B. Bojanov and Q. I. Rahman for stimulating discussions. The author would like to take this opportunity to express his acknowledgent to the INTAS for the financial support (grant INTAS-94-4070 of the European Union).

References

[1] R. Askey, Positive quadrature methods and positive polynomial sums, Approximation Theory 5, C.K. Chui, L. L. Schumaker and J. D. Ward, eds., Academic Press, New York, 1986, 1-30.
[2] K. Atkinson and A. Sharma, a partial characterization of poised Hermite-Birkhoff interpolation problems, Siam J. Numer. Anal. 6 (1969), 230-235
[3] P. Berckmann, *Orthogonal polynomials for engineers and physicists*, The Golem Press, Boulder, Colorado, 1973.
[4] C. Bernardi and Y. Maday, *Approximations spectrales de problèmes aux limites elliptiques*, Spinger-Verlag France, Paris, 1992.
[5] B. Bojanov B. and G. Nikolov, Comparison of Birkhoff type quadrature formulae, Math. Comput., 54(1990), pp. 627–648.
[6] B. Bojanov, G. Grozev and A. A. Zhensykbaev, Generalized Gaussian quadrature formulas for weak Chebychev systems, in Optimal Recovery of Functions, B. Bojanov and H. Wozniakowski, eds., Nova Sciences, New York, 1992, pp. 115-140.
[7] B. Bojanov and A. Guessab, Gaussian quadrature formula of Birkhoff's type, Calcolo 34(1997), pp. 41-50.
[8] B. Bojanov, An inequality of Duffin and Schaeffer type, East J. Approx. 1(1995), pp. 37-46.
[9] B. Bojanov, Total positivity of the spline kernel and its applications, M. Gasca and C. A. Micchelli, eds., Total positivity and its applications, 3-34, Kluwer Academic Publishers, 1996, pp. 3–34.
[10] T. S. Chihara, An introduction to orthogonal polynomials, Math. Appl., vol. 13, Gordon & Breach, New York, 1978.
[11] Ph. J. Davis and Ph. Rabinowitz, *Methods of numerical integration*, Academic Press, London, 1984.
[12] D. K. Dimitrov, On a problem of Turán:(0,2) quadrature formula with a high algebraic degree of precision, Aequationes Math. 41(1991), 230-235.
[13] D. K. Dimitrov, Lacunary quadrature formulae and interpolation singularity, J. Approx. Theory 75(1993), 237-247.
[14] E. A. Van Doorn, Representation and bounds for zeros of orthogonal polynomials and eigenvalues of sign-symmetric tri-diagonal matrices, J. Approx. Theory, v. 51, 1987, pp. 254–266.
[15] S. Elhay and J. Kautsky, IQPACK: Fortran subroutines for the weights of interpolatory quadratures, School of Mathematical Sciences. The Flinders University of South Australia, April 1985.
[16] A. Ezzirani, *Construction de formules de quadrature pour des systèmes de Chebyshev avec applications aux méthodes spectrales*, Thèse de l'Université de Pau, France, 1996.
[17] A. Ezzirani and A. Guessab, A fast algorithm for Gaussian type quadrature formulae with mixed boundary conditions and some lumped mass spectral approximations, to appear in Math. Comp. (1998).
[18] W. Gautschi, A survey of Gauss-Christoffel quadrature formulae, in *E. B. Christoffel,* P. L. Butzer and F. Fehér, eds., Birkhäuser, Basel, 1981, pp. 72-147.
[19] W. Gautschi, Algorithm 726: ORTHPOL— A package of routines for generating orthogonal polynomials and Gauss-type quadrature rules, ACM Trans. Math. Software 20(1994), 21–62.
[20] W. Gautschi, On the construction of Gaussian rules from modified moments. Math. Comp. 24(1970), 245-260.
[21] S. Goedecker, Remark on algorithms to find roots of polynomials. SIAM J. Sci. Computing 15(1994), pp. 1059–1063.
[22] G. H. Golub and J. H. Welsch, Calculation of Gauss quadrature rules, Math. Comp., 23(1969), pp. 221-230.
[23] G. H. Golub, Some modified matrix eigenvalue problems, SIAM Rev., 15(1973), pp. 318-334.
[24] G. H. Golub and J. Kautsky, Calculation of Gauss quadratures with multiple free and fixed knots, Numer. Math., 41(1983), 147-163.

[25] A. Guessab and G. V. Milovanović, An estimate for coefficients of polynomials in L^2 norm, Proc. Amer. Math. Soc., 120(1994), pp. 165-171.
[26] A. Guessab and Q. I. Rahman, Quadrature formulae and polynomial inequalities, J. Approx. Theory, 182(1997), pp. 255-282.
[27] A. Guessab, A fast algorithm to compute Gaussian quadrature formulae for spline function, U.A. CNRS 1204. 9(1998).
[28] S. Karlin and W. J. Studden, *Chebyshev systems: with applications in analysis and statistics*, John Wiley & sons, inc. 1966.
[29] M. G. Krein, "The ideas of P. L. Chebyshev and A. A. Markov in the theory of limiting values of integrals and their further developments, Uspekhi Fiz. Nauk, 1951, 3-120 (Russian); Amer. Math. Soc. Transl. Ser. 2, 12 (1951), pp. 1-122.
[30] B. N. Parlet, The symmetric eigenvalue problem, Prentice-Hall, Englewood Cliffs, 1980.
[31] G. Labelle, Concerning polynomials on the unit interval, Trans. Amer. Math. Soc., (1968), pp. 321-326.
[32] C. A. Micchelli and T. J. Rivlin, Numerical integration rules near Gaussian formulas, Israel J. Math., v. 16(1973), pp. 267–299.
[33] C. A. Micchelli and T. J. Rivlin, Quadrature formulae and Hermite-Birkhoff interpolation, Advances in Mathematics, 11(1973), pp. 93–112.
[34] G. V. Milovanović, Construction of s-orthogonal polynomials and Turán quadratures, in Approx. Theory III, Niš, 1987, G. V. Milovanović, ed. Univ. Niš, 1988, pp. 311-328.
[35] G. V. Milovanović, D. S. Mitronović and Th. M. Rassias, *Topics in polynomials: Extremal problems, inequalities and zeros*, World Scientific Publishing, 1994.
[36] F. Peherstorfer, Characterization of positive quadrature formulae, SIAM J. Math. Anal., 12(1981), pp. 935–942.
[37] F. Peherstorfer, Characterization of positive quadrature formulae II, Siam. J. Math. Anal., 15(1984), pp. 1021–1030.
[38] F. Peherstorfer, On positive quadrature formulas, in Numerical integration IV, International Series of Numerical Mathematics, H. Brass and G. Hämmerlin, eds., Birkhäuser Verlag Basel, 112(1993), pp. 297–313.
[39] A. Ronveaux and W. V. Assche, Upward extension on the Jacobi matrix for orthogonal polynomials, J. approx. Theory, 86(1996), 335-357.
[40] H. J. Schmid, A note on positive quadrature rules, Rocky Mountain J. Math., v. 19, 1989, pp. 395–404.
[41] D. D. Stancu and A. H. Stroud, Quadrature formulas with simple and multiple fixed nodes. Math. Comp., 17(1963), 384-394.
[42] G. Szegö, *Orthogonal Polynomials*, Colloquium Publication, v. 23, 4th ed., Amer. Math. Soc., Providence, R. I., 1975.
[43] P. Turán, On some open problems of approximation theory, J. Approx. Theory 29(1980), 23-85.
[44] D. S. Watkins, Some perspectives on the eigenvalue problem, Siam Review, 35(1993), pp. 430-471.
[45] Y. Xu, Quasi-orthogonal polynomials, quadrature, and interpolation, J. Math. Anal. Appl., 182(1994), pp. 779–799.
[46] Y. Xu, A characterization of positive quadrature formulae, Math. Comp., 62(1994), pp. 703–718.

Current address: Department of applied mathematics, University of Pau, 64000, Pau, France.
E-mail address: allal.guessab@univ-pau.fr

JOURNAL OF CONCRETE AND APPLICABLE MATHEMATICS,VOL.3,NO.2,169-185,2005,COPYRIGHT 2005 EUDOXUS PRESS,LLC

A Cyclic Subgraph Methodology for Estimating de Bruijn Weight Class Distributions

Gregory L. Mayhew
Boeing Phantom Works
P. O. Box 516
MC S306 – 5145
St Louis, MO 63166-0516
tel: 314 – 232 – 0667
email: gregory.l.mayhew@boeing.com

Abstract

Order n de Bruijn sequences are the period 2^n binary sequences produced by an n stage feedback shift register. At present, determining the weight class distribution of the gen-erator functions is theoretically and computationally intractable for all but the smallest n. This paper combines graphical, combinatorial, and numerical techniques to develop a method, illustrated by n=7, that estimates an unknown de Bruijn weight class distribution.

1. Introduction

The *order n de Bruijn sequences* are the $2^{2^{n-1}-n}$ period 2^n binary sequences in which every binary n-tuple occurs exactly once. The de Bruijn sequences are equivalent to Hamiltonian circuits through the de Bruijn graph. The de Bruijn sequences can be algebraically constructed by n-stage nonlinear feedback shift registers and categorized by the Hamming weight of the truth tables of the generating functions [1]. The complete weight class distributions for de Bruijn sequences are known for $n \leq 6$ [2]. For $n \geq 7$, the extreme weight classes can be calculated by counting rooted trees of the PCR and CCR decompositions of the state vector space V(n) [1]. Since a similar algorithm is not known yet for the interior weight classes, limited information has been obtained by exhaustive computer searches [2, 3].

The de Bruijn graph contains many cyclic subgraphs which have equivalent representations as short cycles in the feedback function $x_n \oplus g(x_{n-1} \ldots x_2\, x_1)$ [4, 5]. These short cycles are restrictions on the odd weight truth tables for any given order n. Since exhaustive testing of truth tables is not feasible currently for most orders n, subsets of truth table states can be tested for the existence of short cycles. Convolving results from disjoint subsets produces an upper bound on each weight class in an order n distribution. The numerical and combinatorial effects of these short cycle restrictions greatly improve the bound over the combinatorial bound

obtained by counting all the cycle decompositions possible for any odd weight truth table. The underlying weight class distribution can be estimated by combining the cyclic subgraph bound with the symmetry groups [6] in an order n and the ratio of remaining truth tables to de Bruijn sequences.

2. Weight Classes

The weight w of $x_n \oplus g(x_{n-1} \ldots x_2\ x_1)$ is the number of logical ones (Hamming weight) among the 2^{n-1} entries in the truth table of $g(x_{n-1} \ldots x_2\ x_1)$. Truth tables which produce de Bruijn sequences have odd weight between the minimum weight $Z(n)-1$ and the maximum weight $2^{n-1}-Z^*(n)+1$, inclusive [1]. $Z(n)$ and $Z^*(n)$ are the number of cycles from the pure and complementing cycling registers, respectively.

3. Combinatorial Upper Bound

Let C(j,r) denote the binomial coefficient j choose r [7]. For odd weight feedback functions which produce de Bruijn sequences, the values at g(0) and $g(2^{n-1}-1)$ are 1. Each order n odd weight class w has $C(2^{n-1}-2,w-2)$ truth table candidates. Using order n = 7 to illustrate that this is a weak upper bound, $C(62,w-2) > 2^{57}$ for all odd w in $27 \le w \le 39$, which means that this naive upper bound would allow placement of all order 7 de Bruijn sequences into any one of several odd weight classes. However empirical evidence indicates that all odd weight classes in the proper weight range are nonempty and that the overall de Bruijn weight class distribution has a Gaussian shape [2]. The Cyclic Subgraph Methodology uses short cycles to significantly reduce this combinatorial upper bound.

4. Ratios

Collectively for an order $n \ge 7$, the ratio of truth table candidates to de Bruijn sequences approaches 2^{n-3}. At n = 7, the total number of $C(2^{n-1}-2,w-2)$ truth table candidates is 2,305,741,118,628,006,748 and the total number of $2^{2^{n-1}-n}$ de Bruijn sequences is 144,115,188,075,855,872. The order 7 candidate to sequence ratio is 15.999292992. Within the Cyclic Subgraph Methodology, the quality or "goodness" metric is the reduction of this 2^{n-3} ratio.

Individually by weight class w within an order n, this 2^{n-3} ratio is closely approximated only at the middle weights $2^{n-2}-1$, $2^{n-2}+1$, and $2^{n-2}+3$. The candidate to sequence ratio is significantly

higher than 2^{n-3} at the minimum weight Z(n)-1 and the maximum weight 2^{n-1}-Z*(n)+1. Nevertheless, a first approximation to an order n weight class distribution is $C(2^{n-1}-2,w-2) / 2^{n-3}$ for each valid weight w. Later, this approximation will be compared with the results from the Cyclic Subgraph Methodology.

5. Cyclic Subgraphs and Short Cycles

Most feedback functions with odd weight produce cycles with lengths $k < 2^n$. These short cycles are cyclic subgraphs of the de Bruijn graph as shown in Figures 1 and 2. Let $\beta(n,k)$ denote the number of distinct length k cycles producible in an n stage shift register using the feedback function $x_n \oplus g(x_{n-1} \ldots x_2\ x_1)$.

The distinct and canonical forms of the length k cycles are the cyclic equivalency classes (CEC) in a Bounded Synchronization Code [1]. The number of CECs is

$$\beta(k,k) = \frac{1}{k}\sum_{d|k}\mu(d)2^{k/d} = \frac{1}{k}\sum_{d|k}\mu\left(\frac{k}{d}\right)2^k .$$

For each length k cycle, there exists some shift register length m that is the shortest shift register which can produce this particular length k cycle and guarantee each state has a unique successor. This value m is the minimum span of the length k cycle. A length k cycle and its minimum span is called an [m,k] sequence [5]. (Note m is assigned here instead of the usual n to distinguish the minimum span m from the order n de Bruijn shift register.)

An [m,k] sequence is a sequence of least period k in which all successive k sets of m adjacent digits (called ‘m windows’) are distinct. If $m \le n$, then the length k cycle is producible in an n stage shift register with a unique successor for each state. If m > n, then the length k cycle is *not* producible in an n stage shift register with a unique successor for each state. For example, the length 7 cycle (0101111) is a [4,7] cycle but *not* a [3,7] cycle. As will be explained later, when applied to a particular order n, Cyclic Subgraph Methodology uses [m,k] sequences with $m \le n$.

Figure 1 illustrates the following cyclic subgraphs for an n=4 stage shift register: the short cycle (1) has length k=1, has minimum span m=1, and is a [1,1] sequence; the short cycle (01) has length k=2, has minimum span m=2, and is a [1,2] sequence; the short cycle (011) has length k=3, has minimum span m=2, and is a [2,3] sequence; the short cycle (0001) has length k=4, has minimum span m=3, and is a [3,4] sequence; and the short cycle (00111) has length k=5, has minimum span m=3, and is a [3,5] sequence.

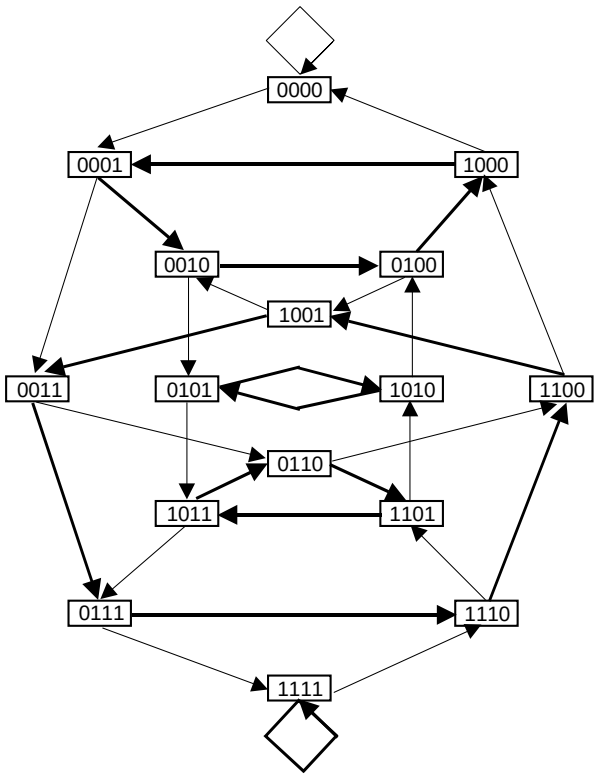


Figure 1: Cycles with Lengths 1 through 5

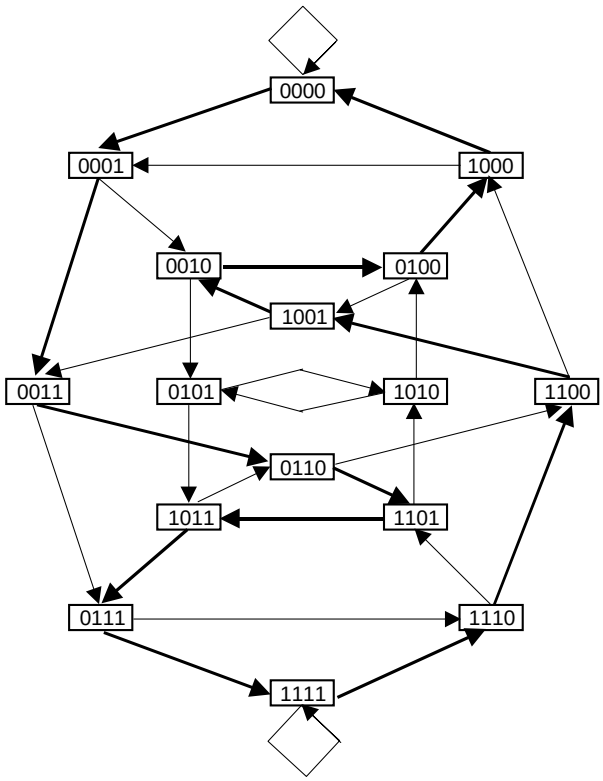


Figure 2: Cycle with Length 14

Table I
[m,k] Sequences for $k \leq 8$

Cycle length k	Minimum span m	[m,k] Sequences
1	1	0, 1
2	1	01
3	2	001, 011
4	2	0011
	3	0001, 0111
5	3	00011, 00111
	4	00001, 00101, 01011, 01111
6	3	000111, 001011, 001101
	4	000011, 000101, 001111, 010111
	5	000001, 011111
7	3	0001011, 0001101, 0010111, 0011101
	4	0000101, 0000111, 0001111, 0101111
	5	0000011, 0010011, 0011011, 0011111
	6	0000001, 0001001, 0101011, 0110111 0010101, 0111111
8	3	00010111, 00011101
	4	00001011, 00001101, 00001111, 00010011, 00011001, 00101101, 00101111, 00110111, 00111011, 00111101
	5	00000101, 00000111, 00011011, 00011111, 00100111, 00101011, 00110101, 01011111
	6	00000011, 00001001, 00010101, 00111111, 01010111, 01101111
	7	00000001, 01011011, 00100101, 01111111

Figure 2 illustrates the cyclic subgraph for an n=4 stage shift register with the short cycle (00001101111001) that has length k=14, has minimum span m=4, and is a [4,14] sequence.

Table I categorizes the CECs for $k \leq 8$ into [m,k] sequences using the lexographically least member of each CEC. Any truth table $x_n \oplus g(x_{n-1} \ldots x_2\ x_1)$ which produces an [m,k] sequence cannot produce a de Bruijn sequence. In a subsequent section, these [m,k] sequences become the basis of tests for length k short cycles and these tests expedite the identification of truth tables that cannot produce de Bruijn sequences.

6. Enumerating β(n,k)

Observe β(n,k) is the total number of [m,k] sequences with length k and miniminum span $m \leq n$. To enumberate β(n,k), first construct all cycles with length k, second test the minimum span of each length k cycle, and third tally all length k cycles with $m \leq n$.

The testing for minimum span proceeds directly from the notion of 'm windows'. Let $(a_0, a_1, ..., a_{k-1})$ be a sequence having least period k from a CEC. By definition, $(a_0, a_1, ..., a_{k-1})$ is an [m,k] sequence only if there does not exist any i and j, $0 \leq i < k$, $0 < j < k$, for which the (i+1)th 'm window' equals the (i+j+1)th 'm window', where indices are computed mod k. This 'm window' testing process is easily programmed on a computer.

Table II summarizes the combinations of n and k at which formulas for β(n,k) are known [4, 5]. The function $\phi_{k,r}$ is defined as the number of integers $t \leq k$ satisfying $(k,t) \leq r$, where (k,t) denotes the greatest common divisor of k and t. The applicable domains for these formulas are illustrated in Figure 3.

Table III extends previous results by enumerating β(n,k) in the unsolved region labelled D in Table II and in Figure 3. An arrow indicates that the value perpetuates in that row or column. When the Cyclic Subgraph Methodology is applied to a particular order n, β(n,k) represents the maximum number of tests for short cycles at each length k.

7. Short Cycle Testing

The [m,k] sequences become the basis for tests which expedite the identification of truth tables that cannot produce de Bruijn sequences. By this method alone, testing truth tables for subcycles up to one-third the sequence length is required to verify a particular truth table produces a de Bruijn sequence. In practice, the majority of invalid truth tables can be quickly eliminated with some subcycle testing and the remaining truth tables tested by attempting to construct a full period sequence from the particular truth table.

Theorem: If $m \leq n$, then a cycle of length k with minimum span m can be structured into a logical test for an n stage shift register that eliminates truth tables which contain that particular length k cycle.

Table II
Formulas for β(n,k)

Region	Domain	β(n,k)
A	n > k	$\beta(n,k) = \beta(k,k)$
B	n = k	$\beta(n,k) = \beta(k,k)$
C	0 ≤ s = k-n ≤ 1/4 k+1	$\beta(n,k) = \beta(k,k) - 2^{s-2}\phi_{k,s-1} - \sum_{t=1}^{s-2}\sum_{j=0}^{s-t-2}\sum_{2\le q<1+(s-d_t-j)/t}\mu(q)2^{d_i+e_j}$
	n = k - 1	$\beta(n,k) = \beta(k,k)$
	n = k - 2	$\beta(n,k) = \beta(k,k) - \phi_k$
	n = k - 3	$\beta(n,k) = \beta(k,k) - 2\phi_{k,2} + 2$
	n = k - 4	$\beta(n,k) = \beta(k,k) - 4\phi_{k,3} - 2(k,2) + 10$
	n = k - 5	$\beta(n,k) = \beta(k,k) - 8\phi_{k,4} - (k,3) + 19$
	n = k - 6	$\beta(n,k) = \beta(k,k) - 16\phi_{k,5} - 4(k,2) - 2(k,3) + 48$
D		UNSOLVED
E	$k = 2^n$	$\beta(n,k) = 2^{2^{n-1}-n}$ de Bruijn sequences
F	$k > 2^n$	$\beta(n,k) = 0$

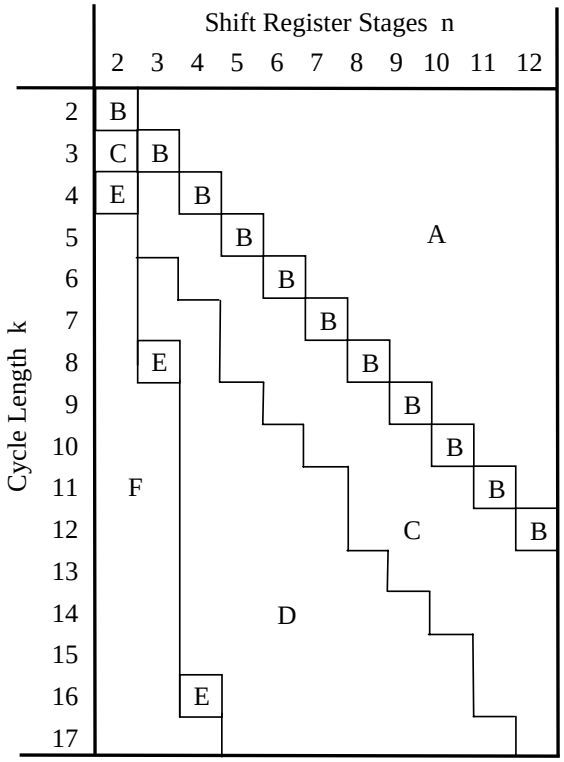


Figure 3: Domains for β(n,k) Formulas

Table III
Values of $\beta(n,k)$ for $n \le 12$ and $k \le 17$

k \ n	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**β(k,k)**
2	1	1	→									1
3	2	2	→									2
4	1	3	3	→								3
5	0	2	6	6	→							6
6	0	3	7	9	9	→						9
7	↓	4	8	12	18	18	→					18
8		2	12	20	26	30	30	→				30
9		0	14	32	46	50	56	56	→			56
10		0	17	57	73	85	95	99	99	→		99
11		↓	14	78	124	154	168	176	186	186	→	186
12			13	113	217	271	309	325	331	335	335	335
13			12	154	348	482	552	590	608	618	630	630
14			20	208	574	877	1009	1083	1119	1139	1155	1161
15			32	300	944	1502	1826	1996	2102	2142	2168	2182
16			16	406	1528	2638	3370	3718	3894	3986	4038	4080
17			0	538	2456	4618	6066	6872	7282	7496	7600	7710

Proof: Let $(a_0, a_1, ..., a_{k-1})$ be a sequence having least period k from a CEC. Length k subcycles are constructed in an order n shift register as follows. The methods are slightly different for $k < n$ and $n \le k < 2^n$, where k is the cycle length and n is the register length. For $k < n$, the fundamental sequence is extended cyclically until its new length $k' > n$. Once $n < k$ or $k' < 2^n$, the 'n windows' test is the following. For $0 \le j \le k-1$ construct k 'n+1 windows', $(a_j, a_{j+1}, ..., a_{j+n-1}, a_{j+n})$, where indices are computed modulo k. Note that $(a_j, a_{j+1}, ..., a_{j+n-1})$ represents the truth table state $(x_n\ x_{n-1}\ ...\ x_2\ x_1)$ and (a_{j+n}) represents the value of $x_n \oplus g(x_{n-1}\ ...\ x_2\ x_1)$ at that state. Because the bottom half of the truth table is the complement of the top half for a de Bruijn sequence, all conditions on $x_n \oplus g(x_{n-1}\ ...\ x_2\ x_1)$ can be translated to conditions on $g(x_{n-1}\ ...\ x_2\ x_1)$. A valid test exists on $g(x_{n-1}\ ...\ x_2\ x_1)$ for that k cycle if four conditions are

satisfied. First, all 'n windows' are distinct because $m \leq n$. Second, if the all zeros state is part of the k cycle, then $g(0)=1$. Third, if the all ones state is part of the k cycle, then $g(2^{n-1}-1)=1$. Fourth, translation of truth table state assignments from the bottom half to the top half does not create any conflicting assignments for $g(x_{n-1} \ldots x_2\ x_1)$ at some state because $m \leq n$.

As examples, consider length k cycles in a n=7 stage shift register. The length 2 cycle (01) exists if $g(21)=1$ and $g(42)=1$. The length 3 cycle (001) exists if $g(9)=1$ and $g(18)=0$ and $g(36)=1$. Similarly, the length 4 cycle (0111) exists if $g(29)=0$ and $g(46)=0$ and $g(55)=1$ and $g(59)=1$. Likewise, the length 9 cycle (000100101) exists if $g(4)=0$ and $g(9)=0$ and $g(10)=1$ and $g(17)=1$ and $g(18)=1$ and $g(20)=0$ and $g(34)=0$ and $g(37)=0$ and $g(40)=1$.

Note that different length k cycles can place different conditions on some subset of states. Again, as an example, consider length k cycles in a n=7 stage shift register. The length 3 cycle (001) requires $g(9)=1$ and $g(18)=0$ whereas the length 9 cycle (000100101) requires $g(9)=0$ and $g(18)=1$.

Theorem: If a feedback function $g(x_{n-1} \ldots x_2\ x_1)$ with odd weight w, $Z(n)-1 \leq$ odd weight $\leq 2^{n-1}-Z^*(n)+1$, does not produce any length k cycle for $1 \leq k \leq 2^n/3$, then $g(x_{n-1} \ldots x_2\ x_1)$ produces a de Bruijn sequence.

Proof: For $n > 2$, the number of cycles into which the n dimensional vector space V(n) is decomposed by a shift register is even or odd according to whether weight w is even or odd [1]. All truth tables producing de Bruin sequences have odd weight. So if one cycle is not produced, then three is the least number of subcycles possible. Any subcycle longer than $2^n/3$ also produces a subcycle shorter than $2^n/3$, so subcycle testing up to $2^n/3$ determines whether or not the truth table produces a period 2^n sequence.

8. Cyclic Subgraph Methodology

The exact and complete weight class distribution for any order $n \geq 7$ could be determined by producing all feedback functions $g(x_{n-1} \ldots x_2\ x_1)$ with odd weight, testing these feedback functions for the $\beta(n,k)$ cycles with length $k \leq 2^n/3$, and then counting only those feedback functions which produce full period sequences. Testing for all short cycles with lengths $2 \leq k \leq 2^n/3$ requires variable access across all states in $g(x_{n-1} \ldots x_2\ x_1)$. However, this approach currently is not feasible due to combinatorial size with readily available computer resources.

On the other hand, partitioning the states into feasible combinatorial sizes, testing these disjoint partitions for short cycles with lengths $2 \le k \le 2^n/3$, and then convolving the results is an approach that is currently feasible. Some short cycle tests may not be included because they could have states in more than one partition. Also, the lengths of cycles tested could be less than $2^n/3$ because the partitions may not have sufficient states to support the longer cycles. As a result, this alternate approach produces an upper bound rather than counting exactly the number of de Bruijn sequences in any or all weight classes for that order n.

The Cyclic Subgraph Methodology for a particular order n has these steps.

1) Determine the cardinality s of the largest set of states for which all 2^s conditions in $g(x_{n-1} \ldots x_2\ x_1)$ can be examined for length k cycles.
2) Set length k=2. Initialize all partitions P_i to the null set of states from $g(x_{n-1} \ldots x_2\ x_1)$.
3) Select an unassigned short cycle with length k from the [m,k] sequences with $m \le n$. If all short cycles of length k with $m \le n$ have been considered, then increment k while $2 \le k \le 2^n/3$ and return to the start of step 3.
4) Determine the test conditions on states in function $g(x_{n-1} \ldots x_2\ x_1)$ for the existence this length k cycle. Valid conditions are produced only when the [m,k] sequence has $m \le n$.
5) The states used in the test for that length k cycle are assigned to some partition $\{P_0, P_1, \ldots, P_j\}$ of states of $g(x_{n-1} \ldots x_2\ x_1)$ provided these conditions are satisfied:
 i) if all of the states do not exist in any partition, create a new partition.
 ii) if some of the states exist in one partition and the other states do not exist in any partition, add the unassigned states to the existing partition provided the maximum partition size s is not exceeded (otherwise discard the length k cycle).
 iii) if some of the states exist in one partition and the other states exist in another partition, merge the two partitions provided the maximum partition size s is not exceeded (otherwise discard the length k cycle).
6) Return to step 3 until short cycles with increasing lengths cannot be included due to the constraints on partitions (e.g., size, disjoint).
7) For each partition $\{P_0, P_1, \ldots, P_j\}$, build all 2^p binary patterns on the $p = |P_i|$ states in partition P_i. Test each pattern for the short cycles which defined that partition. Patterns with short cycles are eliminated. Count the remaining patterns by weight.

8) The overall result, the count of the remaining patterns by total weight w, is obtained by combining the partial results from two partitions at a time until the results of all partitions $\{P_0, P_1, ..., P_j\}$ are merged. Suppose two partitions P_t and P_u contain t and u states, respectively. The count of surviving patterns of weight r with $0 \le r \le w$ is obtained by combining two partitions at a time using the Vandermonde Identity [7]

$$\binom{t+u}{r} = \sum_{i=0}^{r} \binom{t}{i}\binom{u}{r-i}$$

9) Determine the performance metric for the results. For an individual weight class w, compare the final results with $C(2^{n-1}-2,w-2)$. For a weight class distribution for order n, sum the final results for all weight classes, divide this sum by the number of order n de Bruijn sequences, and compare the quotient with 2^{n-3}.

Note that step 3 gives preference to the shortest available subcycle since intuitively a shorter subcycle eliminates more invalid truth tables than a longer subcycle. Furthermore, step 7 eliminates truth tables which produces short cycles and counts the survivors. Invariably, if there is more than one partition and step 7 does not use all cycles up to length $2^n/3$, then the results are an upper bound rather than the exact number of de Bruijn sequences. Also, if a particular partition was determined by a small number of k cycles, then the number of surviving binary patterns on the states could be counted analytically using inclusion / exclusion [7] rather than by exhaustive testing.

9. Order 7 Cyclic Subgraph Bound

Order 7 is the smallest order for which the complete weight class distribution is unknown. Order 7 also has some partial results available for comparative purposes so this order illustrates the Cyclic Subgraph Methodology application.

The computations were performed using a Sun Ultra 60 workstation which contains a Superscalar Ultra SPARC-II version 9 CPU operating at 300 MHz, 2 Megabyte cache, 4 Gigabyte of RAM, and Solaris 6.2 operating system. The numerical algorithms were written in the C programming language.

Given these computer resources, the limit on partition size was 36 states. This was borne out in the CPU time profile discussed later. The initial partitions were determined by the length 2, 3, 4, and 5 cycles. All possible combinations of merging these partitions were considered

based on including the maximum number of length 6 and 7 cycles while not exceeding 36 states total. This resulted in two partitions labelled River and Ocean (for convenience) as shown in Table IV. Note that the length 6 cycles (001011), (001101), and (000011) are not included because they each have states in both partitions.

Table IV
Summary of Order 7 Partitions

Name	**States in $g(x_6x_5x_4x_3x_2x_1)$**	**Short Cycles $k \leq 7$ in Partition**
River (26 states)	1, 2, 3, 4, 5, 8, 9, 10, 11, 16, 17, 18, 19, 20, 21, 32, 33, 34, 36, 37, 40, 41, 42, 48, 50, 52,	01, 001, 0001, 00101, 00001, 000101, 000001, 0000101, 0000001, 0001001, 0010101
Ocean (36 states)	6, 7, 12, 13, 14, 15, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 35, 38, 39, 43, 44, 45, 46, 47, 49, 51, 53, 54, 55, 56, 57, 58, 59, 60, 61, 62	011, 0011, 0111, 00011, 00111, 01011, 01111, 000111, 001111, 010111, 011111, 0001111, 0101111, 0011011, 0011111, 0110111

Cycles up to length 25 were included as possible in the appropriate partition. Table V summarizes the number of short cycle tests included in each partition, where $\beta(n,k)$ defines the maximum number of cycles which could have been included at any length k.

All 2^{26} and 2^{36} binary patterns on the states in the River and Ocean partitions, respectively, were constructed, examined for short cycles, and tallied by weight if they survived. The Ocean partition was tested up to length 10, 13, 16, and 25 cycles in order to investigate the point of diminishing returns for including longer cycle lengths. The CPU hours required and the percent of remaining Ocean patterns are shown in Table VI. The Ocean program execution time in CPU hours exhibited a profile given by

$$\binom{36}{r}\left(3\cdot 10^{-9}\right)3^{(k-10)/3}$$

where r is the pattern weight and k is the maximum length subcycle tested.

Table V
Short Cycle Test Summary for Cyclic Subgraph Upper Bound of Order 7 de Bruijn Weight Class Distribution

Short Cycle Length k	**β(7,k)**	**Total Cycles Tested**	**River Cycles Tested**	**Ocean Cycles Tested**
2	1	1	1	0
3	2	2	1	1
4	3	3	1	2
5	6	6	2	4
6	9	6	2	4
7	18	9	4	5
8	30	12	4	8
9	50	17	5	12
10	85	22	5	17
11	154	30	7	23
12	271	44	9	35
13	482	65	13	52
14	877	97	18	79
15	1,502	116	14	102
16	2,638	152	21	131
17	4,618	209	13	196
18	8,105	292	16	276
19	14,262	396	12	384
20	24,931	499	14	485
21	43,912	668	26	642
22	76.236	861	20	841
23	132,632	1,164	36	1,128
24	229,990	1,446	16	1,430
25	397,260	1,700	0	1,700

Table VI
Ocean Partition Profile

Maximum Length Cycle k	**CPU Hours**	**Percent Patterns Remaining**
10	206	55.734032
13	620	53.047629
16	1,860	51.866302
25	50,145	50.959738

The River and Ocean results were combined for each weight class w using

$$\sum_{i=0}^{w-2} {}_{26}River_i \;\; {}_{36}Ocean_{w-2-i}$$

(dual of Vandermonde Identity) to yield an upper bound for that weight class as shown in Table VII. Summing all weight class bounds and dividing by the 2^{57} sequences yields an overall ratio of 4.077159482. The Cyclic Subgraph Methodology metric is a 74.52 percent reduction of the 15.999292992 metric from the original combinatorial based upper bound on weight class sizes.

Table VII
Upper Bounds on Weight Class Distribution
of Order 7 de Bruijn Sequences

Weight Class	**Naive Upper Bound C(62,w-2)**	**Cyclic Subgraph Upper Bound D(7,w)**
19	739,632,519,584,070	98,666,900,393,138
21	4,282,083,008,118,300	774,674,197,918,760
23	18,412,956,934,908,690	4,017,542,518,481,138
25	59,678,358,445,158,600	14,542,557,734,155,318
27	147,405,545,359,541,742	38,106,646,712,313,044
29	279,692,573,246,309,972	74,084,027,212,710,224
31	409,894,288,378,212,890	108,641,336,612,781,434
33	465,428,353,255,261,088	121,460,568,740,676,744
35	409,894,288,378,212,890	104,147,518,340,124,508
37	279,692,573,246,309,972	68,634,482,648,537,036
39	147,405,545,359,541,742	34,714,653,123,132,692
41	59,678,358,445,158,600	13,413,012,490,444,814
43	18,412,956,934,908,690	3,927,199,071,108,794
45	4,282,083,008,118,300	861,077,229,690,120
47	739,632,519,584,070	139,077,193,488,948
49	93,052,749,919,920	16,177,749,516,244
51	8,308,281,242,850	1,313,493,786,292
53	508,271,323,092	71,175,705,612
55	20,286,591,270	2,406,502,980
Total	2,305,741,118,628,006,748	587,580,605,551,467,840
Ratio	15.999292992	4.077159482

10. Order 7 Weight Class Distribution

A combinatorial based approximation to an order n de Bruijn weight class distribution is given by $C(2^{n-1}-2,w-2) / 2^{n-3}$ for each odd weight w, $Z(n)-1 \le w \le 2^{n-1}-Z^*(n)+1$. This is each weight class binomial coefficient size divided by the overall ratio of the sum of the weight class sizes to the number of de Bruijn sequences.

Using the Cyclic Subgraph Methodology, the underlying order n weight class distribution is estimated by combining the calculated upper bound with the symmetry groups in an order n and the ratio of remaining truth tables to de Bruijn sequences.

Let D(n,w) be the upper bound of the weight class w which results from the Cyclic Subgraph Methodology. Let R(n) be the overall ratio of the sum of the weight class sizes to the number of order n de Bruijn sequences. Let E(n,w) be the estimate of the number of order n de Bruijn sequences in each weight class w. Then R(n) and E(n,w) are

$$R(n) = \sum_{Z(n)-1}^{2^{n-1}-Z^*(n)+1} D(n,w) \Big/ 2^{2^{n-1}-n} \qquad E(n,w) = D(n,w)/R(n).$$

The de Bruijn weight classes within an order n contain large symmetry groups [6]. Thus the estimate for each weight class is conveniently and more concisely expressed in terms of the coefficient in each weight class.

Order 7 is the smallest order for which the complete weight class distribution is unknown. Order 7 also has some partial results available [2, 3] for comparative purposes so this order will be used to illustrate and apply this weight class estimating methodology. In order 7, each weight class has a symmetry group of 2^{26} sequences [2, 6]. The coefficient estimates for the order 7 de Bruijn sequences per weight classes are shown in Table VIII. The first estimate is given by $C(62,w-2) / (15.999292992 \bullet 2^{26})$. The second estimate is $D(7,w) / (4.077159482 \bullet 2^{26})$, where D(7,w) is the appropriate entry from Table VII. Actual coefficients are listed where known.

In all order 7 weight classes for which the actual coefficient is known, the Cyclic Subgraph Methodology estimate is much closer than the combinatorial based value. Presumably, the Cyclic Subgraph Methodology estimate at weight class 19 would have been even closer if more length 7 cycle tests could have been included. The Cyclic Subgraph Methodology estimates also suggest that the middle weight classes $2^{n-2}-1$ and $2^{n-2}+1$ contain more order 7 de Bruijn sequences than suggested by the combinatorial bound and 2^{n-3} ratio.

Table VIII
Coefficient Estimates for Weight Class
Distribution of Order 7 de Bruijn Sequences

Weight Class	Coefficient from C(62,w-2) Bound	Coefficient from D(7,w) Bound	Actual Coefficient [2, 3]
19	688,867	360,607	91,125
21	3,988,177	2,831,271	
23	17,149,160	14,683,274	
25	55,582,259	53,149,993	
27	137,288,178	139,271,787	
29	260,495,518	270,761,553	
31	381,760,673	397,061,257	
33	433,483,087	443,912,857	
35	381,760,673	380,637,296	
37	260,495,518	250,844,612	
39	137,288,178	126,874,763	
41	55,582,259	49,021,742	
43	17,149,160	14,353,087	
45	3,988,177	3,147,056	
47	688,867	508,298	
49	86,666	59,126	52,240
51	7,738	4,800	3,773
53	474	260	160
55	19	9	3
Total	2,147,483,648	2,147,483,648	

In the two order 7 weight classes for which intermediate projections are available, the Cyclic Subgraph Methodology estimate also is closer that the combinatorial derived estimate. A computer program written in C is currently being used to attempt an “exhaustive search” of order 7 de Bruijn weight classes 21 and 47. Weight classes 21 and 47 have 4,282 trillion and 739 trillion truth table candidate cases, respectively, for an unknown number of sequences. From these searches, the weight class 21 and 47 estimates are $1{,}259{,}021 * 2^{26}$ sequences and $526{,}578 * 2^{26}$ sequences, respectively. As seen in Table VIII, both of these estimates are more consistent with the estimates from the Cyclic Subgraph Methodology than the estimates from the combinatorial upper bound method. Finally, empirical evidence again indicates that the valid order 7 weight classes are nonempty and the overall order 7 weight class distribution has a Gaussian shape (albeit truncated on the low weight side).

11. Conclusions

At present, the weight class distributions of the de Bruijn sequence generator functions are theoretically unresolved and computationally intractable for all but the smallest orders n. As a benchmark, a computer program was constructed for an exhaustive search of order 7 de Bruijn weight classes 21 and 47. With the same workstation resources, this program examines 1 billion cases per CPU hour. At this rate, completely examining the order 7 weight classes would require 2.3 billion CPU hours.

As an alternative to exact weight class distributions which are currently unattainable, graphical, numerical, and combinatorial techniques were combined to develop a method for estimating an unknown de Bruijn weight class distribution. With just 50 thousand CPU hours, the Cyclic Subgraph Methodology produced a higher fidelity estimate than the combinatorial derived approach as measured by the known partial order 7 de Bruijn weight class distribution. Perhaps access to a specialized processor such as a hypercube might lead to more data points on the order 7 weight class distribution and provide an opportunity to further assess the accuracy of the estimates based on the Cyclic Subgraph Methodology presented herein.

12. References

[1] H. Fredricksen, “A survey of full length nonlinear shift register cycle algorithms”, SIAM Review 24, 1982, pp. 195-229.

[2] G. L. Mayhew, “Weight class distributions of de Bruijn sequences”, Discrete Mathematics, Vol. 126, 1994, pp. 425-429.

[3] G. L. Mayhew, “Further results on de Bruijn weight classes”, Discrete Mathematics, Vol. 232, 2001, pp. 171-173.

[4] P. Bryant and J. Christensen, “The enumeration of shift register sequences”, Journal Combinatorial Theory, Series A, Vol. 35, 1983, pp. 154-172.

[5] Z. Wan, R. Xiong, and M. Yu, “On the number of cycles of short length in the de Bruijn Good Graph G_n”, Discrete Mathematics, Vol 62, 1986, pp. 85-98.

[6] E. R. Hauge and J. Mykkeltveit, “The analysis of de Bruijn sequences of non-extremal weight”, Discrete Mathematics, Vol 189, 1998, pp. 133-147.

[7] D. Cohen, Basic Techniques of Combinatorial Theory, John Wiley & Sons, New York, 1978.

JOURNAL OF CONCRETE AND APPLICABLE MATHEMATICS,VOL.3,NO.2,187-197,2005,COPYRIGHT 2005 EUDOXUS PRESS,LLC

A Positive Linear Operator for the Approximation of Functions of Two Variables

Alexandra Ciupa

Technical University of Cluj-Napoca, Romania

e-mail: ciupa.alexandra@math.utcluj.ro

Abstract. We consider a positive linear operator $P_{m,n}$ defined on the space of functions of two variables having exponential growth at infinity. We study the properties of approximation of a function f by means of its image $P_{m,n}f$ and we estimate the order of approximation.

1 Introduction

In a recent paper [2], using a method given by A. Jakimovski and D. Leviatan [3], we introduced a generalized Szasz type operator. Let us remind it and some of its properties.

We denote the weight function by $w_p(x) = e^{-px}$, $x \geq 0$, where $p > 0$ and we consider $C_p = \{f \in C[0,\infty) : \ w_p f$ is uniform continuous and bounded on $[0,\infty)\}$, with the norm

$$\|f\|_p = \sup_{x\in[0,\infty)} w_p(x)|f(x)|.$$

We define the positive linear operators

$$P_n(f;x) = \frac{1}{\mathrm{ch1ch}nx} \sum_{k=0}^{\infty} p_{2k}(nx) f\left(\frac{2k}{n}\right),$$

where $f \in C_p$, $\mathrm{ch}x$ is the hyperbolic cosine of x and p_{2k} are the polynomials defined by the relation

$$\mathrm{ch}u\mathrm{ch}ux = \sum_{k=0}^{\infty} p_{2k}(x)u^{2k}, \tag{1}$$

that is

$$p_{2k}(x) = \frac{(1+x)^{2k} + (1-x)^{2k}}{2(2k)!}.$$

We will use some properties of operator P_n.

Lemma A. *If $x \in [0, \infty)$ and $n \in \mathbb{N}$, we have*

$$P_n(e_0; x) = 1$$

$$P_n(e_1; x) = x\mathrm{th}nx + \frac{1}{n}\mathrm{th}1$$

$$P_n(e_2; x) = x^2 + (1 + 2\mathrm{th}1)\frac{x}{n}\mathrm{th}nx + \frac{1}{n^2}(1 + \mathrm{th}1)$$

and

$$P_n((t-x)^2; x) \leq \frac{3}{n}(x+1),$$

where $e_i(x) = x^i$, $i \in \{0, 1, 2\}$ and $\mathrm{th}x$ is the hyperbolic tangent of x.

Lemma B. *Let $p > 0$, $r > p$ and n_0 be a natural number such that $n_0 > \dfrac{p}{\ln r - \ln p}$. For all $x \in [0, \infty)$ and $n \geq n_0$ we have*

$$w_r(x)P_n(e^{pt}; x) \leq 2\frac{\mathrm{ch}\left(e^{p/n}\right)}{\mathrm{ch}1}$$

and

$$w_r(x)P_n((t-x)^2 e^{pt}; x) \leq A_{p,r}\frac{\mathrm{ch}\left(e^{p/n}\right)}{\mathrm{ch}1} \cdot \frac{x+1}{n},$$

where $A_{p,r}$ is a positive constant depending only on p and r.

Lemma C. *If $f \in C_p$ and r, p and n_0 are numbers like in Lemma B, then*

$$\|P_n f\|_r \leq 2\|f\|_p \frac{\mathrm{ch}(e^p)}{\mathrm{ch}1}.$$

2 Definition of operators in two variables

Now we consider an extension of operator P_n to the case of two variables.

We denote by $C(\mathbb{R}_+^2)$ the space of real valued functions of two variables, continuous on $\mathbb{R}_+^2 = \{(x,y) : \ x \geq 0, \ y \geq 0\}$. We will use the weight function $w_p(x) = e^{-px}$, $x \geq 0$, $p > 0$ and we will consider for $p, q > 0$ and $(x,y) \in \mathbb{R}_+^2$ the weight

$$w_{p,q}(x,y) = w_p(x) w_q(y).$$

Let $C_{p,q} = \{f \in C(\mathbb{R}_+^2) : \ w_{p,q} f \text{ is uniform continuous and bounded on } \mathbb{R}_+^2\}$, with the norm

$$\|f\|_{p,q} = \sup_{(x,y)\in\mathbb{R}_+^2} w_{p,q}(x,y)|f(x,y)|. \tag{2}$$

For each $f \in C_{p,q}$ we consider the following positive linear operator

$$P_{m,n}(f;x,y) = \frac{1}{\mathrm{ch}^2 1 \mathrm{ch} mx \mathrm{ch} ny} \sum_{j=0}^{\infty}\sum_{k=0}^{\infty} p_{2j}(mx) p_{2k}(ny) f\left(\frac{2j}{m}, \frac{2k}{n}\right), \tag{3}$$

where $m, n \in \mathbb{N}$, $(x,y) \in \mathbb{R}_+^2$.

Remark 1. If p_{2k} are defined by the relation

$$\mathrm{ch} ux = \sum_{k=0}^{\infty} p_{2k}(x) u^{2k},$$

we obtain

$$p_{2k}(x) = \frac{x^{2k}}{(2k)!},$$

and we get to the operators considered by M. Lesniewicz and L. Rempulska [4]

$$L_{m,n}(f;x,y) = \frac{1}{\mathrm{ch} mx \mathrm{ch} ny} \sum_{j=0}^{\infty}\sum_{k=0}^{\infty} \frac{(mx)^{2j}}{(2j)!} \cdot \frac{(ny)^{2k}}{(2k)!} f\left(\frac{2j}{m}, \frac{2k}{n}\right).$$

Remark 2. If $f(x,y) = f_1(x) f_2(y)$ with $f_1 \in C_p$, $f_2 \in C_q$, $p > 0$, $q > 0$, then for all $(x,y) \in \mathbb{R}_+^2$ and $m, n \in \mathbb{N}$, the following holds

$$P_{m,n}(f;x,y) = P_m(f_1;x) P_n(f_2;y). \tag{4}$$

3 Some properties of $P_{m,n}$ operators

In order to study the approximation properties of $P_{m,n}$ operators we need some auxiliary results. First we study the effect of the operators on the test functions.

Lemma 1. *If $(x,y) \in \mathbb{R}^2_+$ and $m,n \in \mathbb{N}$ the following holds:*

$$P_{m,n}(e_{0,0};x,y) = 1$$

$$P_{m,n}(e_{1,0};x,y) = x\text{th}mx + \frac{1}{m}\text{th}1$$

$$P_{m,n}(e_{0,1};x,y) = y\text{th}ny + \frac{1}{n}\text{th}1$$

$$P_{m,n}(e_{2,2};x,y) = x^2 + y^2 + (1+2\text{th}1)\left(\frac{x}{m}\text{th}mx + \frac{y}{n}\text{th}ny\right) + \left(\frac{1}{m^2} + \frac{1}{n^2}\right)(1+\text{th}1)$$

where $e_{0,0}(t,\tau) = 1$, $e_{1,0}(t,\tau) = t$, $e_{0,1}(t,\tau) = \tau$ and $e_{2,2}(t,\tau) = t^2 + \tau^2$.

Proof. The easiest way to prove these relations is to use remark 2 and Lemma A.

We can also obtain these results by calculus. For instance

$$P_{m,n}(e_{1,0};x,y) = \frac{1}{\text{ch}^2 1 \text{ch}mx\text{ch}ny}\sum_{k=0}^{\infty} p_{2k}(ny)\sum_{j=0}^{\infty} p_{2j}(mx)\frac{2j}{m}.$$

From (1) we have:

$$\sum_{k=0}^{\infty} p_{2k}(ny) = \text{ch}1\text{ch}ny.$$

By using the derivative of (1) with respect to the variable u and replacing u by 1 and x by mx it results

$$\sum_{j=0}^{\infty} p_{2j}(mx)\frac{2j}{m} = \frac{1}{m}\text{sh}1\text{ch}mx + x\text{ch}1\text{sh}mx.$$

Hence

$$P_{m,n}(e_{1,0};x,y) = x\text{th}mx + \frac{1}{m}\text{th}1.$$

Lemma 2. *Let $p,q > 0$, $r > p$, $s > q$ be fixed numbers. We consider m_0, n_0 natural numbers such that*

$$m_0 > \frac{p}{\ln r - \ln p} \quad \textit{and} \quad n_0 > \frac{q}{\ln s - \ln q}.$$

Then for all $m \geq m_0$ and $n \geq n_0$ the following holds

$$\|P_{m,n}\|_{r,s} \leq 4\|f\|_{p,q}\frac{\mathrm{ch}(e^p)\mathrm{ch}(e^q)}{\mathrm{ch}^2 1}.$$

Proof. For all $(x,y) \in \mathbb{R}_+^2$, we have

$$w_{r,s}(x,y)|P_{m,n}(x,y)| \leq$$

$$\leq w_{r,s}(x,y)\frac{1}{\mathrm{ch}^2 1 \mathrm{ch}mx \mathrm{ch}ny}\sum_{j,k=0}^{\infty} p_{2j}(mx)p_{2k}(ny)\left|f\left(\frac{2j}{m},\frac{2k}{n}\right)\right|.$$

According to relation (2), we have

$$\left|f\left(\frac{2j}{m},\frac{2k}{n}\right)\right| \leq \|f\|_{p,q}e^{p\frac{2j}{m}}e^{q\frac{2k}{n}},$$

hence

$$w_{r,s}(x,y)|P_{m,n}(x,y)| \leq w_{r,s}(x,y)\frac{1}{\mathrm{ch}^2 1 \mathrm{ch}mx \mathrm{ch}ny}\cdot$$

$$\cdot\|f\|_{p,q}\sum_{j,k=0}^{\infty} p_{2j}(mx)p_{2k}(ny)e^{p\frac{2j}{m}}e^{q\frac{2k}{n}} =$$

$$= w_r(x)w_s(y)\|f\|_{p,q}P_m(e^{pt};x)P_n(e^{q\tau};y) \leq$$

$$\leq \|f\|_{p,q}\|P_m e^p\|_r\|P_n e^q\|_s.$$

Now we can use Lemma C and we get to the desired inequality.

Theorem 1. *Let $p,q>0$ and $r>p$, $s>q$ be fixed numbers.*

If $f \in C_{p,q}^1$ and m_0, n_0 are natural numbers such that

$$m_0 > \frac{p}{\ln r - \ln p} \quad \text{and} \quad n_0 > \frac{q}{\ln s - \ln q},$$

then there is a positive constant $K_{p,q,r,s}$ depending on p,q,r,s such that

$$w_{r,s}(x,y)|P_{m,n}(f;x,y) - f(x,y)| \leq$$

$$\leq K_{p,q,r,s}\left(\|f'_x\|_{p,q}\sqrt{\frac{x+1}{m}}+\|f'_y\|_{p,q}\sqrt{\frac{y+1}{n}}\right).$$

Proof. Let $(x,y)\in\mathbb{R}^2_+$ be a fixed point. For $f\in C^1_{p,q}$ and all $(t,\tau)\in\mathbb{R}^2_+$ we have

$$f(t,\tau)-f(x,y)=\int_x^t f'_u(u,\tau)du+\int_y^\tau f'_v(x,v)dv$$

and, for $m,n\in\mathbb{N}$,

$$P_{m,n}(f;x,y)-f(x,y)=P_{m,n}\left(\int_x^t f'_u(u,\tau)du;x,y\right)+$$

$$+P_{m,n}\left(\int_y^\tau f'_v(x,v)dv;x,y\right).$$

We multiply this relation by the weight function and we have

$$w_{r,s}(x,y)|P_{m,n}(f;x,y)-f(x,y)|\leq w_{r,s}P_{m,n}\left(\int_x^t f'_u(u,\tau)du;x,y\right)+$$

$$+w_{r,s}(x,y)P_{m,n}\left(\int_y^\tau f'_v(x,v)dv;x,y\right).$$

According to relation (2) we can write

$$\left|\int_x^t f'_u(u,\tau)du\right|\leq\|f'_x\|_{p,q}\int_x^t e^{pu}e^{q\tau}du\leq$$

$$\leq\|f'_x\|_{p,q}e^{q\tau}(e^{pt}+e^{px})|t-x|.$$

In the same way we obtain

$$\left|\int_y^\tau f'_v(x,v)dv\right|\leq\|f'_y\|_{p,q}e^{pt}(e^{q\tau}+e^{qy})|\tau-y|.$$

Using this inequalities it results:

$$w_{r,s}(x,y)|P_{m,n}(f;x,y)-f(x,y)|\leq$$

$$\leq w_{r,s}\|f'_x\|_{p,q}P_{m,n}(e^{q\tau}(e^{pt}+e^{px})|t-x|;x,y)+$$

$$+w_{r,s}\|f'_y\|_{p,q}P_{m,n}(e^{pt}(e^{q\tau}+e^{qy})|\tau-y|;x,y)=A_{m,n}+B_{m,n}.$$

Next we will use remark 2 and we can write

$$A_{m,n} = e^{-rx}e^{-sy}\|f'_x\|_{p,q}P_n(e^{q\tau};y)\{P_m(e^{pt}|t-x|;x) + e^{px}P_m(|t-x|;x)\}$$

and

$$B_{m,n} = e^{-rx}e^{-sy}\|f'_y\|_{p,q}P_m(e^{pt};x)\{P_n(e^{q\tau}|\tau-y|;y) + e^{qy}P_n(|\tau-y|;y)\}.$$

We will focus on $A_{m,n}$ and we try to estimate it.

We can apply Lemma B and Cauchy's inequality and we obtain:

$$w_s(y)P_n(e^{q\tau};y) \leq 2\frac{\mathrm{ch}\left(e^{q/n}\right)}{\mathrm{ch}1}$$

and

$$w_r(x)P_m(|t-x|e^{pt};x) \leq \sqrt{w_r(x)P_m(e^{pt};x)}\cdot\sqrt{w_r(x)P_m((t-x)^2e^{pt};x)}.$$

Also

$$w_r(x)e^{px}P_m(|t-x|;x) \leq \sqrt{P_m((t-x)^2;x)}\cdot e^{(p-r)x}.$$

By making use of these inequalities and Lemma B we obtain

$$A_{m,n} \leq \|f'_x\|_{p,q}\cdot 2\frac{\mathrm{ch}\left(e^{q/n}\right)}{\mathrm{ch}1}\left\{\sqrt{\frac{2\mathrm{ch}\left(e^{p/m}\right)}{\mathrm{ch}1}}\cdot\sqrt{A_{p,r}\frac{\mathrm{ch}\left(e^{p/m}\right)}{\mathrm{ch}1}}\cdot\frac{x+1}{m} + \right.$$

$$\left. + e^{(p-r)x}\sqrt{\frac{3(x+1)}{m}}\right\} = \|f'_x\|_{p,q}K_{p,r}\sqrt{\frac{x+1}{m}}.$$

In the same way we obtain

$$B_{m,n} \leq \|f'_y\|_{p,q}K_{q,s}\sqrt{\frac{y+1}{n}}.$$

By adding $A_{m,n}$ and $B_{m,n}$ we get the desired result.

4 Estimate of the order of approximation

Now we will estimate the order of approximation of function $f \in C_{p,q}$ by operators $P_{m,n}$.

For $u \geq 0$, $v \geq 0$ and $f \in C_{p,q}$ we consider

$$\Delta_{u,v} f(x,y) = f(x+u, y+v) - f(x,y)$$

and the modulus of continuity

$$\omega(f, C_{p,q}; h, \delta) = \sup_{\substack{0 \leq u \leq h \\ 0 \leq v \leq \delta}} \|\Delta_{u,v} f(\cdot,\cdot)\|_{p,q}. \tag{5}$$

Theorem 2. *Let $p > 0$, $q > 0$, $f \in C_{p,q}$ and let $r > p$, $s > q$ be fixed numbers. If m_0, n_0 are fixed numbers such that*

$$m_0 > \frac{p}{\ln r - \ln p} \quad \text{and} \quad n_0 > \frac{q}{\ln s - \ln q},$$

then there is a positive constant K depending only on p, q, r, s such that

$$w_{r,s}(x,y)|P_{m,n}(f;x,y) - f(x,y)| \leq K\omega\left(f, C_{p,q}; \sqrt{\frac{x+1}{m}}, \sqrt{\frac{y+1}{n}}\right)$$

holds for all $(x,y) \in \mathbb{R}^2_+$ and $m \geq m_0$, $n \geq n_0$.

Proof. We consider the Steklov mean of $f \in C_{p,q}$

$$f_{h,\delta} = \frac{1}{h\delta} \int_0^h \int_0^\delta f(x+u, y+v) du\, dv$$

where $h, \delta > 0$ and $(x,y) \in \mathbb{R}^2_+$.

We have

$$f_{h,\delta}(x,y) - f(x,y) = \frac{1}{h\delta} \int_0^h \int_0^\delta [f(x+u, y+v) - f(x,y)] du\, dv =$$

$$= \frac{1}{h\delta} \int_0^h \int_0^\delta \Delta_{u,v} f(x,y) du\, dv.$$

Because

$$\frac{\partial f_{h,\delta}}{\partial x}(x,y) = \frac{1}{h\delta} \int_0^\delta [f(x+h, y+v) - f(x, y+v)] dv$$

and

$$\frac{\partial f_{h,\delta}}{\partial y}(x,y) = \frac{1}{h\delta} \int_0^h [f(x+u, y+\delta) - f(x+u, y)] du$$

it results that $f_{h,\delta} \in C^1_{p,q}$.

By making use of relations (2) and (5) we can write successively:

$$\|f_{h,\delta} - f\|_{p,q} = \sup_{(x,y)\in\mathbb{R}^2_+} w_{p,q}(x,y)|f_{h,\delta}(x,y) - f(x,y)| = \tag{6}$$

$$= \sup_{(x,y)\in\mathbb{R}^2_+} w_{p,q}(x,y)\frac{1}{h\delta}\left|\int_0^h\int_0^\delta \Delta_{u,v}f(x,y)du\,dv\right| \leq$$

$$\leq \frac{1}{h\delta}\int_0^h\int_0^\delta \sup_{(x,y)\in\mathbb{R}^2_+} w_{p,q}(x,y)|\Delta_{u,v}f(x,y)|du\,dv =$$

$$= \frac{1}{h\delta}\int_0^h\int_0^\delta \|\Delta_{u,v}f\|_{p,q}du\,dv \leq$$

$$\leq \frac{1}{h\delta}\int_0^h\int_0^\delta \sup_{\substack{0\leq u\leq h\\ 0\leq v\leq\delta}} \|\Delta_{u,v}f(\cdot,\cdot)\|_{p,q}du\,dv = \omega(f, C_{p,q}; h, \delta).$$

Next we can proceed in the same way:

$$\left\|\frac{\partial f_{h,\delta}}{\partial x}\right\|_{p,q} = \sup_{(x,y)\in\mathbb{R}^2_+} w_{p,q}(x,y)\left|\frac{\partial f_{h,\delta}}{\partial x}(x,y)\right| = \tag{7}$$

$$= \sup_{(x,y)\in\mathbb{R}^2_+} w_{p,q}(x,y)\frac{1}{h\delta}\left|\int_0^\delta [f(x+h,y+v) - f(x,y+v)]dv\right| =$$

$$= \sup_{(x,y)\in\mathbb{R}^2_+} w_{p,q}(x,y)\frac{1}{h\delta}\left|\int_0^\delta [f(x+h,y+v) - f(x,y) + f(x,y) - f(x,y+v)]dv\right| \leq$$

$$\leq \sup_{(x,y)\in\mathbb{R}^2_+} w_{p,q}(x,y)\frac{1}{h\delta}\int_0^\delta (|\Delta_{h,v}f(x,y)| + |\Delta_{0,v}f(x,y)|)dv =$$

$$= \frac{1}{h\delta}\int_0^\delta (\|\Delta_{h,v}f(\cdot,\cdot)\|_{p,q} + \|\Delta_{0,v}f(\cdot,\cdot)\|_{p,q})dv \leq$$

$$\leq \frac{2}{h\delta}\int_0^\delta \sup_{\substack{0\leq u\leq h\\ 0\leq v\leq\delta}} \|\Delta_{u,v}f(\cdot,\cdot)\|_{p,q}dv = \frac{2}{h}\omega(f, C_{p,q}; h, \delta).$$

In a similar manner one obtains:

$$\left\|\frac{\partial f_{h,\delta}}{\partial y}\right\|_{p,q} \leq \frac{2}{\delta}\omega(f, C_{p,q}; h, \delta). \tag{8}$$

For every fixed $(x,y) \in \mathbb{R}^2_+$, $r > p$, $s > q$, $h, \delta > 0$ we have

$$w_{r,s}|P_{m,n}(f;x,y) - f(x,y)| \leq w_{r,s}\{|P_{m,n}(f - f_{h,\delta};x,y)|+$$

$$+|P_{m,n}(f_{h,\delta};x,y)-f_{h,\delta}(x,y)|+|f_{h,\delta}(x,y)-f(x,y)|\}.$$

Using Lemma 2 and (6) for all $m\geq m_0$ and $n\geq n_0$, it results:

$$w_{r,s}|P_{m,n}(f-f_{h,\delta};x,y)|\leq 4\|f-f_{h,\delta}\|_{p,q}\frac{\mathrm{ch}(e^p)\mathrm{ch}(e^q)}{\mathrm{ch}^2 1}\leq$$

$$\leq 4\omega(f,C_{p,q};h,\delta)\frac{\mathrm{ch}(e^p)\mathrm{ch}(e^q)}{\mathrm{ch}^2 1}.$$

Now we apply Theorem 1 and relations (7), (8) and so, for all $m\geq m_0$ and $n\geq n_0$, one obtains:

$$w_{r,s}|P_{m,n}(f_{h,\delta};x,y)-f_{h,\delta}(x,y)|\leq$$

$$\leq K_{p,q,r,s}\left(\left\|\frac{\partial f_{h,\delta}}{\partial x}\right\|_{p,q}\sqrt{\frac{x+1}{m}}+\left\|\frac{\partial f_{h,\delta}}{\partial y}\right\|_{p,q}\sqrt{\frac{y+1}{n}}\right)\leq$$

$$\leq 2K_{p,q,r,s}\omega(f,C_{p,q};h,\delta)\left(\frac{1}{h}\sqrt{\frac{x+1}{m}}+\frac{1}{\delta}\sqrt{\frac{y+1}{n}}\right).$$

Therefore, for all $m\geq m_0$, $n\geq n_0$, $h,\delta>0$ it results that:

$$w_{r,s}(x,y)|P_{m,n}(f;x,y)-f(x,y)|\leq\left[4\frac{\mathrm{ch}(e^p)\mathrm{ch}(e^q)}{\mathrm{ch}^2 1}+\right.$$

$$\left.+2K_{p,q,r,s}\left(\frac{1}{h}\sqrt{\frac{x+1}{m}}+\frac{1}{\delta}\sqrt{\frac{y+1}{n}}\right)+1\right]\omega(f,C_{p,q};h,\delta).$$

Now we choose $h=\sqrt{\frac{x+1}{m}}$, $\delta=\sqrt{\frac{y+1}{n}}$ and we get to the desired result.

From this theorem we can extract a result which we also obtained from Lemma 1.

Corollary. *If $f\in C_{p,q}$, with $p,q>0$, then*

$$\lim_{m,n\to\infty}P_{m,n}(f;x,y)=f(x,y)$$

for all $(x,y)\in\mathbb{R}^2_+$, the convergence being uniform on every rectangle $[0,a]\times[0,b]$, $a>0$, $b>0$.

References

[1] M. Becker, D. Kucharski, R.J. Nessel, *Global approximation theorems for the Szasz-Mirakjan operators in exponential weight spaces*, Linear Spaces and Approximation (Proc. Conf. Oberwolfach, 1977), Birkhäuser Verlag, Basel, ISNM 40(1978), 319-333.

[2] Alexandra Ciupa, *Approximation by a Generalized Szasz Type Operator*, J. Comp. Anal. Appl., vol.5, no.4, October 2003, 413-424.

[3] A. Jakimovski, D. Leviatan, *Generalized Szasz operators for the approximation in the infinite interval*, Mathematica (Cluj), 34, 1969, 97-103.

[4] M. Leśniewicz, L. Rempulska, *Approximation by some operators of Szasz-Mirakjan type in exponential weight spaces*, Glasnik Matematički, Vol.32(52), 1997, 57-69.

Department of Mathematics
Technical University Cluj-Napoca
Str. C. Daicoviciu nr.15
3400 Cluj-Napoca, Romania

JOURNAL OF CONCRETE AND APPLICABLE MATHEMATICS,VOL.3,NO.2,199-219,2005,COPYRIGHT 2005 EUDOXUS PRESS,LLC

On the solution structures of a class of semilinear elliptic equations on R^d

Nedra Belhaj Rhouma
Institut Préparatoire aux Etudes d'Ingénieurs de Tunis.Tunisia

1 INTRODUCTION

In this paper, we consider the semilinear elliptic equation of the form

$$Lu - k_1 f_1(u)u - k_2 f_2(u)u = 0 \text{ in } I\!R^d, d \geq 3, \tag{1}$$

where the operator L is in the following form:

$$L = \sum_{i,j=1}^{d} a_{ij}(x)\frac{\partial^2}{\partial x_i \partial x_j} + \sum_{i=1}^{d} b_i(x)\frac{\partial}{\partial x_i} + c(x)$$

k_1 and k_2 are given Green-tight functions defined on $I\!R^d$, f_1 and f_2 are nonnegative continuous functions.

We shall prove the existence of two types of solutions of (1):

1. Positive solutions u_c of (1) satisfying

$$u_c(x) \to c \text{ as } |x| \to \infty \text{ for some } c \geq 0. \tag{2}$$

2. Maximal solution u_T where

$$u_T = \sup\{u(x) : u \text{ is a positive solution of (1) such that } u \geq T\}.$$

Then we give the asymptotic behaviour of the maximal solution. We show that if $k_2 \cong |x|^{-\nu}$ near ∞ for some $\nu > 2$, , then

$$U_{m_0}(x) = O(|x|^{\frac{\nu-1}{q-1}})$$

when $\sum_{i=1}^{d} |b_i| \neq 0$ and

$$U_{m_0}(x) = O(|x|^{\frac{\nu-2}{q-1}})$$

when $\sum_{i=1}^{d} |b_i| = 0$. Moreover, if $L = \Delta$ and $f_2(y) = y^{q-1}$, $q > 1$, then

$$U_{m_0}(x) \cong (|x|^{\frac{\nu-2}{q-1}})$$

Equation (1) arises from physics and geometry. We give a description of the geometric application of this work (For more details see [5], [6] and [10]).

Let $g = (g_{i,j})_{i,j=1,...,d}$, be a Riemannian metric on $\mathbb{R}^d$ and $k_1(x)$ be its scalar curvature. Let $k_2(x)$ be a given function on $\mathbb{R}^d$. The following question has been raised: Can we find a knew metric g_1 on $\mathbb{R}^d$ such that $k_1(x)$ is the scalar curvature of g_1 and g_1 is conformal to g (that is $g_1 = \Phi g$) for some positive function Φ. When $d \geq 3$, if $\Phi = u^{\frac{4}{d-2}}, u > 0$, then, this is equivalent to the problem of solving the elliptic equation

$$\frac{4(d-1)}{d-2} \triangle_g u - k_1 u - k_2 u^{\frac{d+2}{d-2}} = 0 \text{ in } \mathbb{R}^d$$

where

$$\triangle_g u = \frac{1}{\sqrt{|g|}} \sum_{i,j} \frac{\partial}{\partial x_i} (\sqrt{|g|} g^{i,j} \frac{\partial}{\partial x_j})$$

is the Laplace Beltrami operator, $|g| = \det(g_{i,j}), g^{i,j} = (g_{i,j})^{-1}$.

Equation (1) has been around for a long time. Ni in [10] proved that the equation

$$\triangle u + K(x) u^{\frac{d+2}{d-2}} = 0$$

has infinetely many bounded positive solutions in $\mathbb{R}^d$, when the decay of $|K(x)|$ is faster than $C\,|x|^{-2}$ near ∞ for some positive constant $C > 0$.

Under the same assumptions, this result was improved by Naito in [9] and Li and Ni in [7]. Kenig and Ni in [6] considered the problem

$$Lu - k(x)u + K(x)u^q = 0 \tag{3}$$

in $\mathbb{R}^d$, where $q > 1, Lu = \sum_{i,j=1}^{d} (a_{i,j}(x)u_{x_i})_{x_j}$ and $(a_{i,j}(x))$ is symmetric, positive definite matrix with smooth coefficients of which the eigenvalues are of the order of magnitude $|x|^{a(2-d)}$ at ∞ with $-\infty < a < 1$. they showed that if there exists $\varepsilon > 0$ such that $|K(x)| \leq C_1 |x|^{a(2-d)-2-\varepsilon}$ and $0 \leq k(x) \leq C_2 |x|^{a(2-d)-2-\varepsilon}$ for $|x|$ large and for some positive constants C_1and C_2, then (3) possesses infinitely many bounded positive solutions with the property that each of these solution is bounded below by a positive constant.

Later on, Lin in [8] consider a matrix $(a_{i,j}(x))$ such that $a_{i,j} = a_{j,i}$ for $i, j = 1, ..., d$, are measurable and satisfy the uniform ellipticity condition and a function $w : \mathbb{R}^+ \to \mathbb{R}^+$ locally bounded and such that

$$\int_1^\infty r^{-1} w(r) dr = A < \infty.$$

They showed that there exists a constant $\theta = \theta(d, A)$ such that if $|K(x)| \leq \frac{c_0 w(|x|)}{(1+|x|)^2}$ and $|k(x)| \leq \frac{\theta w(|x|)}{(1+|x|)^2}$ for some positive constant $c_0 > 0$, then for every $c \in (\frac{1}{3}, \frac{1}{2})$, (3) possesses a positive solution u_c satisfying $u_c(x) \to c$ as $|x| \to \infty$.

Chern in [3], consider the problem

$$\triangle u + K_1(x)u^p + K_2(x)u^q = 0 \text{ in } \mathbb{R}^d \tag{4}$$

when $d \geq 3$, $q > p \geq 1$, K_1 and K_2 are two given holder continuous functions on $\mathbb{R}^d$ such that $K_1 \geq 0$, $K_2 \leq 0$ and there exist a Hölder continuous function $\widetilde{K} = \widetilde{K}(|x|)$ and a positive constant T such that

i. $\widetilde{K}(|x|) \leq K_2(x)$,

ii. $\int_0^\infty s\widetilde{K}(s)ds > -\infty$,

iii. $K_1(x) \leq -T^{q-p}.K_2(x)$.

He showed particularly that there exists $T_0 > 0$ such that for all $c > T_0$, the problem (4) has a bounded positive solutions u_c with the property $u_c(x) \to c$ as $|x| \to \infty$. Also he gave some conditions to get the existence of maximal solutions of (4).

2 Preliminary

Let $B(\mathbb{R}^d)$ ($C(\mathbb{R}^d)$ resp.) denote the set of all measurable (continuous resp.) functions on $\mathbb{R}^d$. Given any set A of functions, let A^+ (A_b resp.) be the set of positive (bounded resp.) functions in A.

Let L be a partial differential operaror in the following form:

$$L = \sum_{i,j=1}^{d} a_{ij}(x)\frac{\partial^2}{\partial x_i \partial x_j} + \sum_{i=1}^{d} b_i(x)\frac{\partial}{\partial x_i} + c(x)$$

such that

i. $c \leq 0$,

ii. $a_{ij} = a_{ji}$ for $i, j = 1...n$, are positive measurable and satisfy the uniform ellipticity condition: there exists $\lambda \geq 1$ such that $\sum\limits_{i,j=1}^{d} a_{ij}(x)\xi_i\xi_j \geq \lambda^{-1}|\xi|^2$ for all $x, \xi \in \mathbb{R}^d$.

iii. There exists $\alpha \in]0,1[$ such that

$$\sum_{i,j=1}^{d} |a_{ij}(x) - a_{ij}(y)| + \sum_{i=1}^{d} |b_i(x) - b_i(y)| + |c(x) - c(y)| \leq \lambda |x-y|^{\alpha},$$

iv. $\sum\limits_{i,j=1}^{d} |a_{ij}(x)| + \sum\limits_{i=1}^{d} |b_i(x)| + |c(x)| \leq \lambda$.

We will denote by G^L (G resp.) the Green function of L ($\triangle$ resp.) on $\mathbb{R}^d$. It is known that there exists a constant $C = C(\lambda, \alpha, d)$ such that

$$G^L \leq C \|x-y\|^{2-d}.$$

Let f be a positive measurable function on all $\mathbb{R}^d$ such that

$$x \to \int_{\mathbb{R}^d} G^L(x,y)f(y)(dy)$$

is a continuous bounded potential.

We denote by K^f the mapping defined on $B_b(\mathbb{R}^d)$ as follows: $\forall g \in B_b(\mathbb{R}^d)$

$$K^f(g)(x) = K(fg)(x) = \int_{\mathbb{R}^d} G^L g(y)f(y)(dy).$$

Note that for every $g \in B_b^+(\mathbb{R}^d)$, $K^f(g)$ is a potential on $\mathbb{R}^d$.

In the sequel, for an open subset V of $\mathbb{R}^d$, we will denote by $S(V)$ the set of superharmonic functions on V.

Proposition 2.1 *The bounded operator $I + K^f$ on the banach space $B_b(R^d)$ is invertible and*

$$0 \leq (I + K^f)^{-1}s \leq s \text{ for every } s \in S^+(\mathbb{R}^d)$$

Proof. Similar to the proof of Proposition 2.5 in [2]. ■

Definition 2.1 *A borel function f defined on $\mathbb{R}^d$ is called Green-tight if and only if f satisfies the two conditions:*

$$\lim_{\substack{m(A)\to 0 \\ A\subset R^d}} \left[\sup_{x\in\mathbb{R}^d} \int_A \frac{f(y)(dy)}{\|x-y\|^{d-2}}\right] = 0 \tag{5}$$

and

$$\lim_{M\to\infty} \left[\sup_{x\in\mathbb{R}^d} \int_{y\geq M} \frac{f(y)(dy)}{\|x-y\|^{d-2}}\right] = 0. \tag{6}$$

In the next, we recall the Kato class defined in [1].

Definition 2.2 *A borel function f is said to be in the Kato cass K_d, if and only if*

$$\lim_{\varepsilon\to 0} \left[\sup_{x\in\mathbb{R}^d} \int_{\|x-y\|\leq\varepsilon} \frac{f(y)(dy)}{\|x-y\|^{d-2}}\right] = 0$$

and $f \in K_d^{loc}$, if and only if for every ball B in $\mathbb{R}^d$, $1_B f \in K_d$.

In [11], Zhao introduced the class K_d^∞ as follows

$$K_d^\infty = \left\{ f \in K_d^{loc} : \lim_{M\to\infty} \left[\sup_{x\in I\!R^d} \int_{y\geq M} \frac{f(y)(dy)}{\|x-y\|^{d-2}} \right] = 0 \right\}.$$

Proposition 2.2 *A borel function f defined on $I\!R^d$ is Green-tight if and only if $f \in K_d^\infty$.*

Proof. see [12]. ■

Proposition 2.3 *Let f be a borel function in $I\!R^d$. Suppose that f is in the Kato cass K_d^{loc} and there exits a number $L > 0$ and a positive function φ defined on $[L, \infty[$ such that*

$$\forall\, |x| \geq L, |f(y)| \leq \varphi(|y|) \tag{7}$$

and

$$\int_L^\infty r\varphi(r)dr < \infty.$$

Then f is a Green- tight function in $I\!R^d$.

Proof. Since (5) is obvious, we shall prove that f satisfies the limit equality in (6).

For each $x \in I\!R^d$, we take a coordinate system $(r, \theta_1, ..., \theta_n)$ such that $x = (|x|, 0, ..., 0)$. Then, we have by (7) that $\forall M \geq L$

$$\begin{aligned}
\int_{y\geq M} \frac{|f(y)|}{\|x-y\|^{d-2}} dy &\leq \int_{y\geq M} \frac{\varphi(|y|)}{\|x-y\|^{d-2}} dy \\
&\leq \sigma_{d-2} \int_M^\infty \varphi(r) \left[\int_0^\pi \frac{(\sin\theta_1)^{d-2} d\theta_1}{(r^2 + |x|^2 - 2r|x|\cos\theta_1)^{\frac{d-2}{2}}} \right] r^{d-1} dr \\
&= \sigma_{d-2} \int_M^\infty r\varphi(r) \left[\int_0^\pi \frac{(\sin\theta_1)^{d-2} d\theta_1}{(1 + (\frac{|x|}{r})^2 - 2(\frac{|x|}{r})\cos\theta_1)^{\frac{d-2}{2}}} \right] dr,
\end{aligned}$$

where σ_{d-2} is the area of $S_{d-2}(1)$. On the other hand, we have the estimate, $\forall x \geq 0$

$$\int_0^\pi \frac{(\sin\theta)^{d-2}}{(1 + |x|^2 - 2r|x|\cos\theta)^{\frac{d-2}{2}}} d\theta \leq 2^{d-2}\pi.$$

It follows that

$$\sup_{x\in I\!R^d} \int_{y\geq M} \frac{|f(y)|}{\|x-y\|^{d-2}} dy \leq \sigma_{d-2} 2^{d-2}\pi \int_M^\infty r\varphi(r)dr \to 0 \text{ as } M \to \infty. \blacksquare$$

Remark 2.1 *If* $|f(x)| \leq \frac{c}{\|x\|^{2+\varepsilon}}$ *for* $\|x\|$ *large and for some* $\varepsilon, c > 0$*, then* f *is Green-tight.*

Let $C_0(I\!R^d) = \left\{u \in C(I\!R^d) : \lim_{\|x\|\to\infty} u(x) = 0\right\}$.

Proposition 2.4 *Let* Φ *be a Green-tight function in* $I\!R^d$ *and*

$$F_\Phi = \left\{\varphi \in B(I\!R^d) : |\varphi(x)| \leq |\Phi(x)|, \forall x \in I\!R^d\right\}.$$

Then, the family $\left\{\int_{I\!R^d} \frac{|\varphi(y)|}{\|x-y\|^{d-2}} dy, \varphi \in F_\Phi\right\}$ *is uniformly bounded and equicontinuous in* $C_0(I\!R^d)$*, and consequently it is relatively compact in* $C_0(I\!R^d)$*.*

Proof. For each $\varphi \in F_\Phi$, we have

$$\int_{I\!R^d} \frac{|\varphi(y)|}{\|x-y\|^{d-2}} dy \leq \int_{I\!R^d} \frac{|\Phi(y)|}{\|x-y\|^{d-2}} dy.$$

Hence, the family F_Φ is uniformly bounded and $\lim_{|x|\to\infty} \int_{I\!R^d} \frac{|\varphi(y)|}{\|x-y\|^{d-2}} dy = 0$ uniformly for all function $\varphi \in F_\Phi$.

Finally, let $x, x' \in I\!R^d$, then

$$\left|\int_{I\!R^d} \frac{\varphi(y)}{\|x-y\|^{d-2}} dy - \int_{I\!R^d} \frac{\varphi(y)}{\|x'-y\|^{d-2}} dy\right| \leq \int_{I\!R^d} \left|\frac{1}{\|x-y\|^{d-2}} - \frac{1}{\|x'-y\|^{d-2}}\right| |\Phi(y)|\, dy \to 0$$

as $x \to x'$.

Thus the family F_Φ is equicontinuous. Since it is uniformly bounded, we conclude by Ascoli-Arzela theorem that the family F_Φ is relatively compact in $C_0(I\!R^d)$. ■

Since G and G^L are comparable, then we get the following result.

Corollary 2.1 *Let* Φ *be a Green-tight function in* $I\!R^d$ *and*

$$F_\Phi = \left\{\varphi \in B(I\!R^d) : |\varphi(x)| \leq |\Phi(x)|, \forall x \in I\!R^d\right\}.$$

Then, the family $\left\{\int_{I\!R^d} G^L |\varphi(y)|\, dy, \varphi \in F_\Phi\right\}$ *is uniformly bounded and equicontinuous in* $C_0(I\!R^d)$*, and consequently it is relatively compact in* $C_0(I\!R^d)$*.*

Let $\widetilde{f_1}$ ($\widetilde{f_2}$resp.) denote the map: $y \to f_1(y)y$ ($y \to f_2(y)y$ resp.) and for all $y > 0$, we denote by $\widetilde{k}_y$, the function $x \to \widetilde{k}_y(x) = k_1(x) + yk_2(x)$.

In all the next, we assume that the functions k_1 and k_2 are Green-tight, and that the functions f_1 and f_2 are nonnegative and continuous,

3 Existence of positive solutions u_c for (1)

In this section, we shall prove the existence of positive solutions u_c of (1) which satisfies (2).

We suppose that the following property is satisfied:

(A) $\forall c > 0$, there exists $\alpha_c \geq 0$ such that for every $x \in I\!R^d$, the function

$$y \to -k_1(x)f_1(y)y - k_2(x)f_2(y)y + \alpha_c(|k_2(x)| + |k_1(x)|)y$$

is increasing on $[0, c]$.

Particularly; if f is locally Lipschitz, then *(A)* is satisfied.

In the next, we shall prove the following result:

Theorem 3.1 *We suppose that the set*

$$P = \left\{c > 0 : k_1(x)f_1(c) + k_2(x)f_2(c) \geq 0, \forall x \in I\!R^d\right\}$$

is nonempty.

Then, for all $c \in P$, *there exists a positive solution* u_c *of (1) which satisfies (2).*

Corollary 3.1 *We Suppose that:*

1. $k_2 \geq 0$ *on* R^d,
2. *There exists* $T_0 > 0$ *such that* $\widetilde{k}_{T_0}(x) \geq 0, \forall x \in I\!R^d$,
3. $\lim_{y\to\infty} \frac{f_2(y)}{f_1(y)} = +\infty$.

Then, there exits $m_0 > 0$, *such that for each* $c \geq m_0$, *equation (1) has a positive solution* u_c *which satisfies (2).*

Proof. There exits $m_0 > 0$, such that for each $y \geq m_0$, we have $\frac{f_2(y)}{f_1(y)} \geq T_0$. Hence, for each $c \geq m_0$, we have

$$\begin{aligned} k_1(x)f_1(c) + k_2(x)f_2(c) &= f_1(c)(k_1(x) + k_2(x)\frac{f_2(c)}{f_1(c)}) \\ &\geq f_1(c)\widetilde{k}_{T_0}(x) \geq 0. \end{aligned}$$

The conclusion follows by using Theorem 3.1. ■

Remark 3.1 *Note that the condition k_1 and k_2 are Green-tight is more general than the decay condition $|k_2| \leq cr^{-l}$ near ∞ for some $l > 2$ which appeared first in [9], later in [6] and [10] and it is more general than the condition*

$$\int_0^\infty rK(r)dr < \infty$$

which appeared in [4] and [3].

Moreover, Lin proved in [8] that if $k_2 \geq \frac{C}{\|x\|^{2-\varepsilon}}$, for some $\varepsilon > 0$, then (1) has no positive solution in $\mathbb{R}^d$. He also showed that if $L = \triangle$, $k_2 = 0$ and $k_1 = k_1(r) = \frac{1}{1+r^2}$, equation (1) has no bounded solution in $\mathbb{R}^d$.

In what follows, we shall prove Theorem 3.1.

Definition 3.1 *We say that a function u is a subsolution of (1) if*

$$u + K^{k_1 f_1(u)}(u) + K^{k_2 f_2(u)}(u) \in S(\mathbb{R}^d).$$

We say that a function u is a sursolution of (1) if

$$u + K^{k_1 f_1(u)}(u) + K^{k_2 f_2(u)}(u) \in -S(\mathbb{R}^d).$$

Proposition 3.1 *Suppose that there exists two bounded functions $u_0 \leq v_0$ such that u_0 is a subsolution of (1) satisfyuing (2) and v_0 a sursolution of (1) satisfyuing (2). Then there exits a solution w of (1) which satisfies (2) and such that $u_0 \leq w \leq v_0$.*

Proof. There exit $s \in S(\mathbb{R}^d)$ and $s' \in -S(\mathbb{R}^d)$ such that

$$v_0 + K^{k_1 f_1(v_0)} v_0 + K^{k_2 f_2(v_0)} v_0 = c + s$$

and

$$u_0 + K^{k_1 f_1(u_0)} u_0 + K^{k_2 f_2(u_0)} u_0 = c + s'.$$

Since $\lim_{\|x\|\to\infty} v_0(x) = \lim_{\|x\|\to\infty} u_0(x) = c$, it follows that $\lim_{\|x\|\to\infty} s(x) = \lim_{\|x\|\to\infty} s'(x) = 0$, we get $s \geq 0$ and $-s' \geq 0$ on $\mathbb{R}^d$. Let $c = \max(\|u_0\|, \|v_0\|)$ and α_c satisfying (A_3), then for each $u \in C_b^+(\mathbb{R}^d)$, we set

$$F(u) = -k_1(x)f_1(u)u - k_2(x)f_2(u)u + \alpha_c(|k_2(x)| + |k_1(x)|)u$$

and

$$\widetilde{K} = \alpha_c K^{|k_2|+|k_1|}.$$

Now, we define an operator T on $C_b^+(I\!R^d)$ as follows:

$$\forall u \in C_b^+(I\!R^d), T(u) = (I + \widetilde{K})^{-1}(K^{F(u)} + c).$$

First, we claim that the operator T is increasing. In fact, if $0 \leq u \leq v$, then by Proposition 2.1, $T(v) - T(u) = (I + \widetilde{K})^{-1}K^{F(v)-F(u)} \geq 0$. On the other hand, we have $T(v_0) \leq v_0$. In fact, since $(I + \widetilde{K})(v_0 - T(v_0)) = v_0 + K^{k_1 f_1(v_0)}v_0 + K^{k_2 f_2(v_0)}v_0 - c = s \in S^+(I\!R^d)$, hence, by Proposition 2.1, $v_0 - T(v_0) = (I + \widetilde{K})^{-1}s \geq 0$.

Similarly, we show that $T(u_0) \geq u_0$. Thus, the sequence defined inductively by $u_n = T(u_{n-1})$ and $v_n = T(v_{n-1})$ are monotonous and $u_n \leq v_n$. Let $w = \lim_{n\to\infty} v_n$. Since by Proposition 2.4 the operator T is compact and continuous, we get $T(w) = w$ and consequently $w + K^{k_1 f_1(w)}w + K^{k_2 f_2(w)}w = c$. ■

In what follows, we set $E = \Big\{u \in C_0(I\!R^d) : 0 \leq u \leq c\Big\}$ and we define the operator F as follows:

$$\forall u \in E, F(u) = (I + K^{k_1^+ f_1(u)+k_2^+ f_2(u)})^{-1}c.$$

Lemma 3.1 *$F(E)$ is relatively compact in $C_0(I\!R^d)$.*

Proof. Let $(u_n)_n$ be a sequence in E. Then

$$F(u_n) + K(k_1^+ f_1(u_n) + k_2^+ f_2(u_n)(F(u_n)) = c.$$

By Proposition 2.1, we have

$$0 \leq F(u_n) \leq c.$$

Thus, by Proposition 2.4, the family

$$\Big\{K(k_1^+ f_1(u_n) + k_2^+ f_2(u_n))(F(u_n))) : n \in N\Big\}$$

is relatively compact in $C_0(I\!R^d)$ ■

Lemma 3.2 *The operator F is continuous.*

Proof. Let $(u_n)_n$ be a sequence in E such that $(u_n)_n$ converges uniformly to a function $u \in C_0(\mathbb{R}^d)$. Let us suppose that $(F(u_n))_n$ converges to a function v. Then $v + K^{k_1^+ f_1(u) + k_2^+ f_2(u)}(v) = c$ and therefore $v = F(u)$. ■

Lemma 3.3 *Let $c \geq 0$. Then there exists $v_c \in E$ such that v_c is a subsolution of (1) satisfying (2).*

Proof. We have $F(E) \subset E$. Since F is completely continuous, then by Shauder's fixed point Theorem, we conclude that the operator F has a fixed point v_c satisfying

$$v_c + Kk_1^+ f_1(v_c)v_c + Kk_2^+ f_2(v_c)v_c = c.$$

Hence

$$v_c + Kk_1 f_1(v_c)v_c + Kk_2 f_2(v_c)v_c = c - Kk_1^- f_1(v_c)v_c - Kk_2^- f_2(v_c)v_c.$$

Thus v_c is a subsolution of (1). Since k_1 and k_2 are Green-tight, then

$$\lim_{\|x\|\to\infty} K(|k_1 f_2(v_c)v_c|)(x) = \lim_{\|x\|\to\infty} K(|k_2 f_2(v_c)v_c|)(x) = 0$$

and therefore $\lim_{\|x\|\to\infty} v_c(x) = c$. ■

Now, we are ready to prove Theorem 3.1.

Proof of Theorem 3.1

Since $c \in P$, then $v_0 = c$ is a sursolution of (1) and satisfies (2). Let v_c as in Lemma 3.3, then v_c is a subsolution of (1) which satisfies (2) and $v_c \leq v_0$. Thus, by Proposition 3.1, there exits a solution u_c of (1) such that $v_c \leq u_c \leq v_0$, obviously $\lim_{\|x\|\to\infty} u_c(x) = c$.

4 Existence of maximal solution u_T

For this section, we assume that the following conditions hold:

(F_1) $k_2 \geq 0$ on $\mathbb{R}^d$ and locally bounded,
(F_2) f_1 and f_2 are derivable on $]0, \infty[$,

$(F_3)\lim_{y\to\infty}\frac{\widetilde{f_2}'}{\widetilde{f_1}'}(y) = +\infty$.
There exists $a > 0$ such that :
$(F_4)\widetilde{f_1}$ is increasing on $]a,\infty[$ and $\lim_{y\to\infty}\widetilde{f_1}(y) = \infty$
(F_5) There exist $\delta > 0$ and $q > 1$ such that $f_2(y) \geq \delta y^{q-1}$, for all $y \geq a$.

Example 4.1 *We can choose f_1 and f_2 as follows:*
$f_1 = P$, *where P is a polynome of degree $p \geq 1$.*
$f_1(y) = y^p \log(y)$, *$p \geq 1$.*
$f_2 = Q$, *where Q is a polynome of degree $q > p$.*
$f_2(y) = H(y)\exp(\lambda y)$, *$\lambda > 0$, where H is a polynome.*

Remark 4.1 *For every $x, y > 0$, there exists $c_{x,y} \in]x, y[$ such that*

$$\frac{\widetilde{f_2}(x) - \widetilde{f_2}(y)}{\widetilde{f_1}(x) - \widetilde{f_1}(y)} = \frac{(\widetilde{f_2})'(c_{x,y})}{(\widetilde{f_1})'(c_{x,y})}.$$

Thus, by the assumptions given above, we get that $\lim\frac{\widetilde{f_2}}{\widetilde{f_1}}(y) = +\infty$.

In this section, we shall prove the following results:

Theorem 4.1 *We suppose that there exist a sequence of bounded smooth domains $\{\Omega_i\}_{i\in N}$, a sequence of positive real $(\delta_i)_{i\in I\!N}$ and $T_0 \geq a$ such that*

1. $I\!R^d = \cup_{i=1}^{\infty}\Omega_i$ *with* $\overline{\Omega_i} \subset \Omega_{i+1}$, $i = 1, 2, ...,$
2. $\widetilde{k}_{T_0} \geq 0$ *on* $I\!R^d$,
3. *For each i, $k_2 \geq \delta_i$ on a neighberhood of $\partial\Omega_i$,*
4. *For each i, there exists $1 > \rho_i > 0$ such that $k_1 + (1-\rho_i)k_2T_0 \geq 0$ on a neighberhood of $\partial\Omega_i$,*
5. *There exists a positive solution v of (1) such that $v \geq m_0$, where $m_0 \geq T_0$ satisfies $\frac{(\widetilde{f_2})'(y)}{(\widetilde{f_1})'(y)} \geq T_0$ and $\frac{f_2(y)}{f_1(y)} \geq T_0$ for all $y \geq m_0$.*

Then, the function

$$U_{m_0} = \sup\{u(x) : u \text{ is a positive solution of (1) with } u \geq m_0\}$$

is well defined and is a solution of (1).

Corollary 4.1 *We suppose that*

1. k_1 *and* k_2 *are continuous,*
2. *There exist* $T_0 \geq a$ *and a sequence* $(r_n)_n \to \infty$ *such that* $\widetilde{k}_{T_0}(x) > 0$ *and* $k_2(x) > 0$ *for all* $\|x\| = r_n$*,*
3. $\widetilde{k}_{T_0} \geq 0$ *on* $\mathbb{R}^d$*,*
4. *There exists a positive solution* v *of (1) such that* $v \geq m_0$*, where* $m_0 \geq T_0$ *satisfies* $\frac{(\widetilde{f_2})'(y)}{(\widetilde{f_1})'(y)} \geq T_0$ *and* $\frac{f_2(y)}{f_1(y)} \geq T_0$ *for all* $y \geq m_0$*.*

Then, the function

$$U_{m_0} = \sup\{u(x) : u \text{ is a positive solution of (1) with } u \geq m_0\}$$

is well defined and is a solution of (1).

Proof. $\mathbb{R}^d = \cup_{i=1}^{\infty} B(0, r_i)$. Since k_1 and k_2 are continuous, then for each i, there exist $\delta_i > 0$ and $1 > \rho_i > 0$ such that $\widetilde{k}_{T0} \geq \delta_i$ and $\rho_i k_2 T_0 \leq \delta_i$ on a neighberhood of $B(0, r_i)$. Thus $k_1 + (1 - \rho_i)k_2 T_0 \geq 0$ on a neighberhood of $B(0, r_i)$. The conclusion follows by Theorem 4.1.

■

In what follows we shall prove some preliminary results which will be needed for the proof of Theorem 4.1.

Proposition 4.1 *Let* $\gamma > 0, q > 1$*. Let* $x_0 \in \mathbb{R}^d$*,* $U = \left\{x \in \mathbb{R}^d : \|x - x_0\| < R\right\}$ *for* $R > 0$*. Then there exists a function* u *such that*

$$Lu - \gamma u^q \leq 0 \text{ in } U \tag{8}$$

and

$$\lim_{x \to z} u(x) = \infty, \text{ for all } z \in \partial U.$$

Moreover

$$u(x) = \lambda(R^2 - \|x - x_0\|^2)^{\frac{-2}{q-1}}$$

where $\lambda = C(1 \vee R)^{\frac{3}{q-1}}$ *for a positive constant* C *which depends only on* q,γ, *the dimention* d *and the upper bound for* a_{ij} *and* b_i .

Proof. Let $u = \lambda(R^2 - r^2)^{\frac{-2}{q-1}}$ for some $\lambda > 0$, where $r = \|x - x_0\|$. By a direct computation we get:

$$Lu - \gamma u^q = \lambda(R^2 - r^2)^{\frac{-2q}{q-1}}(c_1 \sum a_{ij} z_i z_j + c_2(R^2 - r^2)(\sum a_{ii} + \sum b_i z_i) + c(R^2 - r^2)^{\frac{2q}{q-1}} - \gamma\lambda^{q-1}) \tag{9}$$

Where $z_i = x_i - x_{0,i}$, $c_1 = 8(q+1)(q-1)^{-1}$, $c_2 = 4(q-1)^{-2}$. Let A and B be the upper bounds $\sum |a_{ii}|$ and $\sum |b_i|$. Then (9) implies (8) if

$$c_1 A r^2 + c_2 A(R^2 - r^2) + B c_2 R^3 - \gamma\lambda^{q-1} \leq 0 \tag{10}$$

for all $0 \leq r \leq R$. The condition (10) holds if

$$\lambda^{q-1} \geq \frac{(c_1 + c_2)AR^2 + Bc_2R^3}{\gamma} \tag{11}$$

which is true for $\lambda = C(1 \vee R)^{\frac{3}{q-1}}$, where $C = (\frac{(c_1+c_2)A+Bc_2}{\gamma})^{\frac{1}{q-1}}$. ■

Notation 4.1 *Let* Ω *be an open subset of* $\mathbb{R}^d$ *and* f *be a measurable function on* $\mathbb{R}^d$, *we denote by*

$$K_\Omega^f = \int_\Omega G_\Omega^L(\cdot, y) f(y) dy$$

where G_Ω^L *denotes the Green function in* Ω.

Proposition 4.2 *Let* $q, \gamma > 0$ *and* u, v *such that* $Lu - \gamma u^q \leq 0$ *and* $Lv - \gamma v^q \geq 0$ *in* Ω. *If* $\liminf_{x \to z} u(x) \geq \limsup_{x \to z} v(x)$ *for every* $z \in \partial\Omega$, *then* $u \geq v$ *on* Ω.

Proof. Let $s = u - v$. Then $\liminf_{x \to z} s(x) \geq 0$ for every $z \in \partial\Omega$ and $Ls - \gamma(\frac{u^q - v^q}{u - v})s \leq 0$. Using the minimum Principle, we obtain $u - v \geq 0$ on Ω. ■

For the next, we consider the problem

$$(\mathrm{E}_m) \quad \begin{cases} Lu - k_1 f_1(u)u - k_2 f_2(u)u = 0 \text{ in } \Omega \\ \lim_{x \to y} u(x) = m, \text{ for all } y \in \partial\Omega \end{cases}$$

Definition 4.1 *We say that u is a subsolution for (E_m) if*

$$\begin{cases} Lu - k_1 f_1(u)u - k_2 f_2(u)u \geq 0 \text{ in } \Omega \\ u \leq m \text{ on } \partial\Omega. \end{cases}$$

We say that u is a sursolution for (E_m) if

$$\begin{cases} Lu - k_1 f_1(u)u - k_2 f_2(u)u \leq 0 \text{ in } \Omega \\ u \geq m \text{ on } \partial\Omega. \end{cases}$$

Proposition 4.3 *We suppose that there exists a subsolution u and a sursolution v for (E_m) such that $u \leq v$, then there exists a solution w of (E_m) such that $u \leq w \leq v$.*

Proof. There exists $s, s' \in S_b(\Omega)$ such that $v + K_\Omega^{k_1 f_1(v) - k_2 f_2(v)}(v) = m + s$ and $u + K_\Omega^{k_1 f_1(u) - k_2 f_2(u)}(u) = m - s'$. Since $v \geq m$ on $\partial\Omega$ and $u \leq m$ on $\partial\Omega$, we conclude that $s \geq 0$ and $s' \leq 0$.

Let $c = Max(\|u\|, \|v\|)$ and α_c satsfiying *(A)*. For each $u \in C_b^+(\mathbb{R}^d)$, we set

$$F(u) = -k_1(x)f_1(u)u - k_2(x)f_2(u)u + \alpha_c(|k_2(x)| + |k_1(x)|)u$$

and

$$\widetilde{K} = \alpha_c K^{|k_2|+|k_1|}.$$

We define an operator T on $C_b^+(\mathbb{R}^d)$ as follows: $T(u) = (I + \widetilde{K})^{-1}(K^{F(u)} + m)$. Hence, as in the proof of the Proposition 3.1, the operator T has a fixed point $u \leq w \leq v$ which is solution of (E_m). ■

Lemma 4.1 *Let Ω be a smooth bounded domain. Suppose that there exists $T_0 \geq a$ such that $\widetilde{k}_{T_0}(x) \geq 0$ on Ω. Let $m_0 \geq T_0$ such that $\frac{f_2(y)}{f_1(y)} \geq T_0$ and $\frac{f_2'(y)}{f_1'(y)} \geq T_0$ for all $y \geq m_0$.*

If the equation (1) possesses a positive solution $v \geq m_0$, then, for each $m \geq m_1 = \max\{v(x) : x \in \partial\Omega\}$, the problem (E_m) has a solution $w_m \geq v$ on Ω. Moreover the sequence $(w_m)_{m \geq m_1}$ is increasing.

Proof. Let $m \geq \max\{v(x) : x \in \partial\Omega\}$. It is obvious that v is a subsolution of (E_m). On the other hand, since $\widetilde{k}_m \geq 0$ for $m \geq m_1$, then m is a sursolution of (E_m). Thus, we conclude by Proposition 4.3 that there exists a solution w_m of (E_m) such that $v \leq w_m \leq m$.

Now, by setting $z = w_{m+1} - w_m$, we get

$$\begin{cases} Lz - hz = 0 \text{ in } \Omega \\ z = 1 \text{ on } \partial\Omega \end{cases}$$

where $h = (k_1 + k_2(\frac{\widetilde{f}_2(w_{m+1}) - \widetilde{f}_2(w_m)}{\widetilde{f}_1(w_{m+1}) - \widetilde{f}_1(w_m)}))\frac{\widetilde{f}_1(w_{m+1}) - \widetilde{f}_1(w_m)}{w_{m+1} - w_m}$. Since $w_{m+1}, w_m \geq m_0$, then $h \geq 0$ on Ω and by the minimum principle, we get $z > 0$ on $\overline{\Omega}$. ■

Proposition 4.4 *Let T_0 and $m_0 \geq T_0$ as in Lemma 4.1. Suppose that there exists a sequence of bounded smooth domains $\{\Omega_i\}_{i\in N}$ and a sequence of positive real $(\delta_i)_{i\in N}$ such that:*

1. *$\mathbb{R}^d = \cup_{i=1}^{\infty}\Omega_i$ with $\overline{\Omega_i} \subset \Omega_{i+1}$, $i = 1, 2, ...,$*
2. *$\widetilde{k}_{T_0} \geq 0$ on $\mathbb{R}^d$,*
3. *For each i, $k_2 \geq \delta_i > 0$ on a neighberhood of $\partial\Omega_i$,*
4. *For each i, there exists $1 > \rho_i > 0$ such that $k_1 + (1 - \rho_i)k_2T_0 \geq 0$ on a neighberhood of $\partial\Omega_i$,*
5. *Equation (1) possesses a positive solution v satisfying $v(x) \geq m_0$ for all $x \in \mathbb{R}^d$.*

Then, there exists a positive solution u_i satisfying

$$\begin{cases} Lu_i - k_1f_1(u_i)u_i - k_2f_2(u_i)u_i = 0 \text{ in } \Omega_i \\ \lim_{x\to y} u_i(x) = \infty \text{ , for all } y \in \partial\Omega_i. \end{cases} \tag{12}$$

Furthermore $u_i \geq u_{i+1} \geq m_0$ on Ω_i for each i.

Proof. Since $\partial\Omega_i$ is smooth, we can use $(x, t\nu(x)), x \in \partial\Omega_i$, $|t| < \varepsilon$, for some $1 > \varepsilon > 0$, as local coordinates for a neigberhood of $\partial\Omega_i$, where $\nu(x)$ is the outward normal of $\partial\Omega_i$ at $x \in \partial\Omega_i$. From the assumptions 3 and 4, we can choose $\varepsilon > 0$ so small that

$$\begin{cases} k_1(x, t\nu(x)) + (1 - \rho_i)k_2(x, t\nu(x))T_0 \geq 0 \\ k_2(x, t\nu(x)) \geq \delta_i > 0 \end{cases}$$

for all $-\varepsilon \leq t \leq 0$ and for all $x \in \partial\Omega_i$. Now, fix $x \in \partial\Omega_i$, $-\varepsilon \leq t < t' < 0$ and let $y = (x, t'\nu(x))$. Put $U_y = \left\{z \in \mathbb{R}^d : \|z - y\| < R\right\}$ where $R = \inf(-t', t' - t)$.

For each i, let $(w_m)_m$ be the sequence defined in Lemma 4.1. Then, on U_y we have

$$\begin{aligned} Lw_m &= k_1 f_1(w_m)w_m + k_2 f_2(w_m)w_m \\ &= (k_1 + (1-\rho_i)k_2\frac{f_2(w_m)}{f_1(w_m)})f_1(w_m)w_m + \rho_i k_2 f_2(w_m)w_m \\ &\geq (k_1 + (1-\rho_i)k_2 T_0)f_1(w_m)w_m + \rho_i k_2 f_2(w_m)w_m \\ &\geq \rho_i \delta \delta_i w_m^q. \end{aligned}$$

By Proposition 4.1, there exists a suitible λ such that the function

$$u(z) = \lambda(R^2 - \|z-y\|^2)^{\frac{-2}{q-1}}$$

satisfies

$$Lu - \rho\delta\delta_i u^q \leq 0 \text{ on } U_y.$$

Consequently, by Proposition 4.2 we get

$$w_m(z) \leq u(z), \text{ for all } z \in U_y.$$

Particularly, for $z = y$, we get

$$w_m(y) \leq u(y) \leq \lambda(R^2)^{\frac{-2}{q-1}}.$$

But, from hypothesis 2, we have that w_m is a subharmonic function on $\mathbb{R}^d$, which yields

$$w_m(y) \leq \lambda(R^2)^{\frac{-2}{q-1}} \quad \forall y \in \Omega_i^t,$$

where $\Omega_i^t = \Omega_i - \{(x, t'\nu(x)), x \in \partial\Omega_i, t < t' < 0\}$. Hence, for every $-\varepsilon \leq t < 0$, $(w_m)_{m\geq m_1}$ is increasing and uniformly bounded in Ω_i^t. Thus $(w_m)_{m\geq m_1}$ converges uniformly in Ω_i^t to a function u_i which satisfies

$$Lu_i - k_1 f_1(u_i)u_i - k_2 f_2(u_i)u_i = 0.$$

Now, let $z = u_i - u_{i+1}$. Then z satisfies

$$Lz - hz = 0 \text{ on } \Omega_i,$$

where $h = (k_1 + k_2(\frac{\widetilde{f}_2(u_i) - \widetilde{f}_2(u_{i+1})}{\widetilde{f}_1(u_i) - \widetilde{f}_1(u_{i+1})}))\frac{\widetilde{f}_1(u_i) - \widetilde{f}_1(u_{i+1})}{u_i - u_{i+1}} \geq 0$. We can choose $\varepsilon > 0$ small enough to get $z > 0$ on $\partial\Omega_i^t$ for $-\varepsilon \leq t < 0$. By using the minimum principle, we get

$$z > 0 \text{ on } \partial\Omega_i^t \text{ for all } \varepsilon \leq t < 0$$

and hence $u_i \geq u_{i+1}$on Ω_i. ■

Next, we give The proof of Theorm 4.1.

Proof of Theorm 4.1.

Consider the sequence of functions $(u_i)_i$ defined in Proposition 4.4. Then, for each i, $u_i \geq v \geq m_0$ in Ω_i. Thus, $(u_i)_i$ converges to a function $\widetilde{u}$ such that $\widetilde{u} \geq m_0$ and $\widetilde{u}$ is a solution of (1). On the other hand, if u is a solution of (1) such that $u \geq m_0$, then, $u \leq u_i$ on Ω_i. Thus, $u \leq \widetilde{u}$ on $\mathbb{R}^d$. This proves that $\widetilde{u} = U_{m_0}$.

5 Asymptotic behaviour of the maximal solution

In this section, we keep the assumptions given in Section 4. We will study the asymptotic behaviour of U_{m_0}. We shall prove the following results.

Theorem 5.1 *Suppose that there exist $T_0 \geq a$ and $1 > \rho > 0$ such that*

1. $k_1 + (1-\rho)k_2\ T_0 \geq 0$ *on* $\mathbb{R}^d$,
2. *There exists a positive solution v of (1) such that $v \geq m_0$, where $m_0 \geq T_0$ satisfies $\frac{f_2(y)}{f_1(y)} \geq T_0$ for all $y \geq m_0$.*
3. $k_2(x) \cong |x|^{-\nu}$ *near* ∞ *for some* $\nu > 2$.

Then, the maximal solution U_{m_0} is such that

1. $U_{m_0}(x) = O(|x|^{\frac{\nu-1}{q-1}})$ *when* $\sum_{i=1}^{d} |b_i| \neq 0$.
2. $U_{m_0}(x) = O(|x|^{\frac{\nu-2}{q-1}})$ *when* $\sum_{i=1}^{d} |b_i| = 0$.

Theorem 5.2 *We suppose that $L = \Delta$ and that the assumptions* 1, 2 *and* 3 *in Theorem 5.1 are satisfied. Moreover, we assume that there exists $\alpha > 0$ such that $k_1 \leq \alpha k_2 T_0$. If $f_2(y) = y^{q-1}$, then, the maximal solution U_{m_0} is such that*

$$U_{m_0}(x) \cong (|x|^{\frac{\nu-2}{q-1}}).$$

Such a result was proved for radial solutions in [4] when $k_1 = 0$, $f_2(y) = y^{q-1}$ and in [3] when $k_1 \leq 0$ and $f_2(y) = y^{q-1}$.

Proof of Theorm 5.1.

Let $C > 0, R_0 > 1$ such that $k_2(r) \geq Cr^{-\nu}$ for all $r \in [R_0, +\infty[$. We fix $x \in I\!R^d$ such that $\|x\| > 2R_0$ and we set $U = B(x, R)$ where $R = \frac{\|x\|}{2}$.

Hence, if v is a solution of (1) such that $v \geq m_0$, we get

$$\begin{aligned} Lv &= k_1 f_1(v)v + k_2 f_2(v)v \\ &= k_1 f_1(v)v + (1-\rho)k_2 f_2(v)v + \rho k_2 f_2(v)v \\ &= \widetilde{f_1}(v)(k_1 + (1-\rho)k_2 \frac{\widetilde{f_2}(v)}{\widetilde{f_1}(v)}) + \rho k_2 \widetilde{f_2}(v) \\ &\geq \rho\delta k_2 v^q. \end{aligned}$$

Thus, for each $y \in U$, we have $R_0 \leq \|y\| \leq 3R$ and

$$\begin{aligned} Lv(y) &\geq \rho\delta C \|y\|^{-\nu} v^q(y) \\ &\geq \frac{\rho\delta C}{3^\nu R^\nu} v^q(y). \end{aligned}$$

By Proposition 4.1, the function

$$u(y) = \lambda(R^2 - \|x-y\|^2)^{\frac{-2}{q-1}}$$

satisfies

$$\begin{cases} Lu - \frac{\rho\delta C}{3^\nu R^\nu} u^q \leq 0, \text{ on } U \\ \lim_{y\to z} u(y) = \infty, \ \forall z \in \partial U \end{cases}$$

for

$$\lambda^{q-1} \geq 3^\nu \frac{(c_1+c_2)AR^2 + Bc_2R^3}{\rho\delta C} R^\nu.$$

Thus, if $\sum_{i=1}^{d} |b_i| \neq 0$, we take $\lambda = \lambda_1 = (3^\nu \frac{(c_1+c_2)A+Bc_2}{\rho\delta C})^{\frac{1}{q-1}} R^{\frac{\nu+3}{q-1}}$ and if $\sum_{i=1}^{d} |b_i| = 0$, we take $\lambda = \lambda_2 = (3^\nu \frac{(c_1+c_2)A}{\rho\delta C})^{\frac{1}{q-1}} R^{\frac{\nu+2}{q-1}}$.

In bouth cases, we have

$$v(y) \leq u(y), \forall y \in U$$

and consequently, we get

$$v(x) \leq \lambda (R^2)^{\frac{-2}{q-1}}.$$

Hence, for $\lambda = \lambda_1$, we obtain that

$$v(x) \leq (3^{\nu} \frac{(c_1 + c_2)A + Bc_2}{2^{\nu-1}\rho\delta C})^{\frac{1}{q-1}} \|x\|^{\frac{\nu-1}{q-1}}$$

which yields

$$U_{m_0} = O(\|x\|^{\frac{\nu-1}{q-1}}).$$

For $\lambda = \lambda_2$, we obtain that

$$v(x) \leq (3^{\nu} \frac{(c_1 + c_2)A}{2^{\nu-1}\rho\delta C})^{\frac{1}{q-1}} \|x\|^{\frac{\nu-2}{q-1}}$$

which yields

$$U_{m_0} = O(\|x\|^{\frac{\nu-2}{q-1}}).$$

Proof of Theorm 5.2.

We supoose that $f_2(y) = y^{q-1}$ and $L = \Delta$. Thus, by Theorem 2 in [4], the equation

$$\Delta u - (1 + \alpha) k_2 u^q = 0 \tag{13}$$

has a maximal solution U such that $U(x) \cong |x|^{\frac{\nu-2}{q-1}}$ near ∞. Now, let u_i be the solution of (12), then by setting $s = u_i - U$, we get

$$\Delta s - (1 + \alpha) k_2 \frac{({u_i}^q - U^q)}{u_i - U} s = (k_1 - \alpha k_2 \frac{\widetilde{f}_2(u_i)}{\widetilde{f}_1(u_i)}) \widetilde{f}_1(u_i) \leq 0. \tag{14}$$

Since $s > 0$ on a neigberhood of $\partial\Omega_i$, we get by the maximum principle that $u_i \geq U$ on Ω_i. which yields that $U_{m_0} \geq U$ on Ω_i for all i. Thus by Theorem 5.1 we get $U_{m_0}(x) \cong |x|^{\frac{\nu-2}{q-1}}$ near ∞.

■

References

[1] M.Aizenman and B.Simon: Brownian motion and Harnack inequality for Schrödinger operators. Comm. Pure. Appl. math. 35, 209-273.(1982).

[2] A.Boukricha, W.Hansen and H.Hueber: Continuous solutions of the generalized Schrödinger equation and perturbation of harmonic spaces. Exposition Math. 5. 97-135 (1987).

[3] Chern J L: On the solution structures of the semilinear elliptic equations on $I\!R^d$. Nonlinear. Anal. Method and Appl. Vol 28. N 10. 1741-1750. (1997).

[4] Cheng K.S and W.M Ni., On the structure of the conformal scalar curvature equation on $I\!R^d$. Indiana Univ. Math. J. 41. 261-278. (1992), 493-529.

[5] Kazdan J., prescribing the curvature of a Riemannian manifold, CBMS Regional Conf. Vol.57. Am. Math. Soc. (1985).

[6] Kenig. C. E and Ni. W.M: On exterior Dirichlet problem with applicatios to some nonlinear equations arising in geometry, Am. J. Math. 106. 189-702. (1984).

[7] Y.Li and W.M. Ni: On conformal scalar curvature equations in $I\!R^d$. Duke Math.J.57. 895-924.(1988).

[8] Lin F.H., On the elliptic equation $D_i[a_{i,j}(x)D_j u] - k(x)u + K(x)u^p = 0$, pro. Am. Math. Soc. 95, 219-226. (1985).

[9] Natio M: A note on bounded positive entire solutions of semilinear elliptic equations, Hirochima Math. J. 14 (1984), 211-214.

[10] Ni W.M: On the elliptic equation $\triangle u + k(x)u^{\frac{n+2}{n-2}} = 0$, its generalisation and application in geometry, Indiana. Uni. Math. J. 31, 493-529 (1982).

[11] Zhao Z : Subcriticality, positivity and gaugeability of the Schrödinger operator, Bull. Amer. Math. Soc. 23, 513-517. (1990).

[12] Zhao Z : On the existence of positive solutions of nonlinear elliptic equations.A probabilistic potential theory approach. Duke Math. J. 69,(2). 247-258. (1993).

JOURNAL OF CONCRETE AND APPLICABLE MATHEMATICS,VOL.3,NO.2,221-239,2005,COPYRIGHT 2005 EUDOXUS PRESS,LLC

On the degree of copositive approximation

Ivan V.Smazhenko

Department of Mathematical Analysis, Mechanics-Mathematics Faculty,

Kyiv National University, Kyiv, Ukraine

and

Laboratory ETMA, Faculty of Science and Technics,

University of South, Toulon-Var, Toulon, France.

E-mail address: smazhen@mail.univ.kiev.ua

Abstract

The functions which are continuous and change their sign finitely many times on the segment $[a, b]$ are considered. For such functions all cases are investigated, when the analogs of uniform and pointwise Jackson type estimates are valid for the approximation of function by copositive with it polynomial.

AMS classification: 41A10, 41A25, 41A29.

Key words and phrases: Copositive, polynomial, approximation, Jackson type, pointwise.

1 Introduction

Let $I := [-1, 1]$, and for $s \in \mathbf{N}$ let $Y := \{y_i\}_{i=1}^{s}$, $-1 = y_{s+1} < \cdots < y_1 < y_0 = 1$; $\mathbf{Y}_s$ — the set of all such collections. Finally let $\Delta^{(0)}(Y)$ be the set of continuous functions f on I such that f is non-negative on $[y_i, y_{i-1}]$, when i is odd and is non-positive, when i is

even, and set

$$\Pi(x) := \Pi(x, Y) := \prod_{i=1}^{s}(x - y_i).$$

Note that $f \in \Delta^{(0)}(Y)$ if and only if $f(x)\Pi(x) \geqslant 0,\ x \in I$. A function g is said to be copositive with $f \in \Delta^{(0)}(Y)$ if $g(x)f(x) \geqslant 0,\ x \in I$.

The following notations are used ($k \in \mathbf{N};\ r, n \in \mathbf{N}_0$):

$$C := C(I),$$

$$\|f\| := \max\{|f(x)| : x \in I\},$$

$$C^{(r)} := \{f : f^{(r)} \in C\},$$

$\mathcal{P}_n$ — the space of algebraic polynomials of degree $\leqslant n$,

$$E_n(f) := \inf\{\|f - P_n\| :\ P_n \in \mathcal{P}_n\},$$

$$E_n^{(0)}(f, Y) := \inf\{\|f - P_n\| :\ P_n \in \mathcal{P}_n \cap \Delta^{(0)}(Y)\},$$

$$\rho_n(x) := \frac{1}{n^2} + \frac{\sqrt{1 - x^2}}{n},\ x \in I\ (\rho_0(x) := 1,\ \text{briefly } \rho_n(x) =: \rho),$$

$\omega_k(f; t)$ — the ordinary k-th modulus of smoothness of function $f \in C$.

The main result of this paper is the following theorems.

Theorem 1. *Let $k = 1$, or $r = s - 1$, $k = 2$, or $r \geqslant s$, and $Y \in \mathbf{Y}_s$. If $f \in C^{(r)} \cap \Delta^{(0)}(Y)$, then for every $n \geqslant k + r - 1$ the polynomial $P_n \in \mathcal{P}_n \cap \Delta^{(0)}(Y)$ exists such that*

$$|f(x) - P_n(x)| \leqslant C(k, r, s)\rho_n^r(x)\omega_k(f^{(r)}; \rho_n(x)).$$

Theorem 2. *Let $r \geqslant 1$, or $r = 0$, $k \leqslant 3$, and $Y \in \mathbf{Y}_s$. If $f \in C^{(r)} \cap \Delta^{(0)}(Y)$, then for every $n \geqslant N(Y) \geqslant k + r - 1$ the polynomial $P_n \in \mathcal{P}_n \cap \Delta^{(0)}(Y)$ exists such that*

$$|f(x) - P_n(x)| \leqslant C(k, r, s)\rho_n^r(x)\omega_k(f^{(r)}; \rho_n(x)),$$

and for every $n \geqslant k + r - 1$ the polynomial $P_n \in \mathcal{P}_n \cap \Delta^{(0)}(Y)$ exists such that

$$|f(x) - P_n(x)| \leqslant C(k, r, Y)\rho_n^r(x)\omega_k(f^{(r)}; \rho_n(x)).$$

If we compare Theorems 1 and 2 and Counterexamples from Section 5 (which are valid in the uniform case) we are possible to make some conclusions in graphical form (they are the same in uniform and pointwise case, see Tables I–IV below). In tables I–IV entry $+$ means that at the corresponding k, r, s the estimate for $f \in \Delta^{(0)}(Y)$ and for some polynomial $P_n \in \Delta^{(0)}(Y)$

$$|f(x) - P_n(x)| \leqslant C\rho_n^r(x)\omega_k(f^{(r)}; \rho_n(x)), \quad n \geqslant N \tag{1.1}$$

is valid for $C = C(k, r, s)$ and $N = k + r - 1$. Entry $\oplus$ means that (1.1) is valid for $C = C(k, r, s)$ and $N = N(Y) \geqslant k + r - 1$ and vice versa for $C = C(k, r, Y)$ and $N = k + r - 1$, and it cannot be obtained with $C = C(k, r, s)$ and $N = k + r - 1$. And finally entry $-$ means that (1.1) isn't valid even with the constants which depend on f.

Tables of Truth

I. The case $s = 1$

r	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\cdot^{\cdot^{\cdot}}$
3	$+$	$+$	$+$	$+$	$+$	$\cdots$
2	$+$	$+$	$+$	$+$	$+$	$\cdots$
1	$+$	$+$	$+$	$+$	$+$	$\cdots$
0	$+$	$+$	$\oplus$	$-$	$-$	$\cdots$
	1	2	3	4	5	k

II. The case $s = 2$

r	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\cdot^{\cdot^{\cdot}}$
3	$+$	$+$	$+$	$+$	$+$	$\cdots$
2	$+$	$+$	$+$	$+$	$+$	$\cdots$
1	$+$	$+$	$\oplus$	$\oplus$	$\oplus$	$\cdots$
0	$+$	$\oplus$	$\oplus$	$-$	$-$	$\cdots$
	1	2	3	4	5	k

III. The case $s = 3$

r	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\cdot^{\cdot^{\cdot}}$
3	$+$	$+$	$+$	$+$	$+$	$\cdots$
2	$+$	$+$	$\oplus$	$\oplus$	$\oplus$	$\cdots$
1	$+$	$\oplus$	$\oplus$	$\oplus$	$\oplus$	$\cdots$
0	$+$	$\oplus$	$\oplus$	$-$	$-$	$\cdots$
	1	2	3	4	5	k

IV. The case $s > 3$

$$\begin{array}{cccccccc}
r & \vdots & \vdots & \vdots & \vdots & \vdots & \iddots \\
s & + & + & + & + & + & \cdots \\
s-1 & + & + & \oplus & \oplus & \oplus & \cdots \\
s-2 & + & \oplus & \oplus & \oplus & \oplus & \cdots \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\
1 & + & \oplus & \oplus & \oplus & \oplus & \cdots \\
0 & + & \oplus & \oplus & - & - & \cdots \\
 & 1 & 2 & 3 & 4 & 5 & k
\end{array}$$

Let's give a short history of the problem. The first estimate in copositive approximation of Jackson type was obtained by D.Leviatan[5]. The estimate (1.1) was established with at least one constant depending on Y for $r = 0$, $k \leqslant 3$ by Y.K.Hu, K.A.Kopotun and X.M.Yu[3] and for $r \geqslant 1$ by G.A.Dzyubenko[1] and Y.K.Hu, K.A.Kopotun and X.M.Yu[3] (as the consequence of the result in so-called intertwining approximation). A counterexample for $r = 0$, $k \geqslant 4$ and $s = 1$ was obtained by S.P.Zhou[8]. For more detailed survey one can address to the work by Y.K.Hu and X.M.Yu[4].

We will use the Chebyshev partition of $[-1, 1]$, namely, let $x_{-1} := 1$, $x_{n+1} := -1$ and for each $j = 0,...,n$, set $x_j := x_{j,n} := \cos(j\pi/n)$, $I_j := [x_j, x_{j-1}]$, and $h_j := |I_j| := x_j - x_{j-1}$. The following inequalities take place:

$$\rho_n(x) < h_j < 5\rho_n(x), \qquad x \in I_j,\ j = 1,...,n \tag{1.2}$$

$$h_{j\pm1} < 3h_j, \qquad j = 1,...,n, \tag{1.3}$$

$$\rho_n^2(y) < 4\rho_n(x)(|x-y| + \rho_n(x)), \qquad x, y \in I, \tag{1.4}$$

which implies

$$2(|x-y| + \rho_n(x)) > |x-y| + \rho_n(y), \qquad x, y \in I. \tag{1.5}$$

Let's set for each $j = \overline{1,n}$

$$\Gamma_j(x) := \frac{h_j}{|x - x_j| + h_j}, \qquad \gamma_j(x) := \frac{\rho_n(x)}{\operatorname{dist}(x, I_j) + \rho_n(x)},$$

where $\operatorname{dist}(x, E) = \inf\{|x - y| : y \in E\}$.
We make use of some other inequalities:

$$\sum_{j=1}^{n} h_j \gamma_j^2(x) < 20\rho_n(x), \tag{1.6}$$

$$\gamma_j^2(x) < 16\Gamma_j(x), \quad \Gamma_j^2(x) < 400\gamma_j(x). \tag{1.7}$$

For a given Y, let

$$O_i := (x_{j+1}, x_{j-2}), \qquad \text{if} \quad y_i \in [x_j, x_{j-1}),$$

and set

$$O := \bigcup_{i=1}^{s} O_i.$$

Next let $\widetilde{O}_q := [a_q, b_q]$, $q \leqslant s$, be the connected components of the closure $\overline{O}$ of O, indexed so that $b_{q+1} < a_q$. We will write $j \in H$, if $I_j \cap O = \varnothing$, $j = \overline{1,n}$.
The next inequality is easily verified:

$$\left|\frac{\Pi(x)}{\Pi(y)}\right| \leqslant \left(\frac{|x-y|}{\rho_n(y)} + 1\right)^s, \quad x \in I, \quad y \in I \setminus O. \tag{1.8}$$

2 Auxiliary lemmas

Let $\varphi \in \Phi^k$, i.e., $\varphi(0+) = 0$ and $\varphi(t)$ is nondecreasing while $t^{-k}\varphi(t)$ is nonincreasing on $(0, \infty)$. If function f belongs to $C^{(r)}$ and $\omega_k(f^{(r)}; \cdot) \leqslant \varphi$, we will denote this by $f \in W^r H_k^\varphi$. It is assumed in the paper that $f \in W^r H_k^\varphi \cap \Delta^{(0)}(Y)$ for some k, r and $Y \in \mathbf{Y}_s$. We will stand $c, c_i, i \in \mathbf{N}$ for different constants, which depend only on k, r, s (but not Y).

First we recall one known statement, which we need. It was proved by J.Gilewicz and I.A.Shevchuk[2].

Statement 1. *Let $s \in N$, $Y \in \mathbf{Y}_s$, $k \in N$, $(r+1) \in N$, $[a, b] \subset I$, $h := b - a$, $g \in W^r H_k^\varphi \cap \Delta^{(0)}(Y)$. If* i) $k = 1$ *or* ii) $s = r+1$, $k = 2$ *or* iii) $s \leqslant r$, *then such a polynomial $P_{k+r-1}(x) := P_{k+r-1}(x, g, [a, b], k, r)$ of degree $\leqslant k + r - 1$ exists that for every $x \in [a, b]$*

$$|g(x) - P_{k+r-1}(x)| \leqslant c_*(k, r) h^r \varphi(h),$$

and

$$P_{k+r-1}(x)\Pi(x) \geqslant 0, \qquad x \in [a, b],$$

where $c_*(k, r) =$ const depends only on k and r.

Let's denote by Σ_k the collection of piecewise polynomials (not necessarily continuous) of degree $\leqslant k$ with the knots at the x_j's, and denote by $\Sigma_{k,O}$ the subset of $S \in \Sigma_k$ which are polynomials on each $\widetilde{O}_q$. When we use the notation $S(x_j)$ we mean either $S(x_j+)$ or $S(x_j-)$, when x tends to x_j from the right and from the left respectively. This makes no difference in our arguments. Note that all derivatives of the piecewise polynomial (spline) exist except perhaps at the Chebyshev nodes. When we will use them for Chebyshev nodes, we will make corresponding reservations.

Lemma 1. *If $f \in C[a, b]$ and $f(x) \geqslant 0$ for $x \in [a, b]$, then a polynomial P_{k-1} of degree $\leqslant k - 1$ exists such that*

$$\|f - P_{k-1}\|_{[a,b]} \leqslant c\omega_k(f; h; [a, b]), \tag{2.1}$$

and

$$P_{k-1}(x) \geqslant 0, \quad x \in [a, b], \tag{2.2}$$

where $h = b - a$.

Proof. Let P^*_{k-1} be the polynomial of the best uniform approximation of f in $[a, b]$, and

$$E_k := \|f - P^*_{k-1}\|_{[a,b]}.$$

Let's set

$$P_{k-1}(x) := P^*_{k-1} + E_k.$$

Inequality (2.2) is evident and

$$\|f - P_{k-1}\|_{[a,b]} \leqslant 2E_k.$$

Then (2.1) follows from Whitney's inequality. □

An obvious consequence of Statement 1, Lemma 1 and (1.3) is following

Lemma 2. *Let $r \geqslant s$, or $s = r + 1$, $k = 2$, or $k = 1$. Then there exists $S \in \Sigma_{k+r-1,O} \cap \Delta^{(0)}(Y)$ such that*

$$\|f - S\|_{I_j} \leqslant ch_j^r \omega_k(f^{(r)}; h_j), \quad j = 1, ..., n. \tag{2.3}$$

Lemma 2 can be formulated also for many other k, r, s, provided we take $n > N(Y)$, where $N(Y)$ is taken so that for each $n > N(Y)$ every $\widetilde{O}_q$ contains exactly one point $y_q \in Y$. So we have

Lemma 3. *Let $r \geqslant 1$, or $r = 0$, $k = 1, 2$ and suppose that $n > N(Y)$. Then a piecewise polynomial $S \in \Sigma_{k+r-1,O} \cap \Delta^{(0)}(Y)$ exists such that (2.3) holds.*

Copositive piecewise polynomial, which gives proper estimate, can be constructed also when $r = 0$, $k = 3$. For this case we have

Lemma 4. *Let $r = 0$, $k = 3$ and $n > N_1(Y)$. There is an $S \in \Sigma_{2,O} \cap \Delta^{(0)}(Y)$ such that*

$$\|f - S\|_{I_j} \leqslant c\omega_3(f; h_j), \quad j = 1,..., n. \tag{2.4}$$

Proof. We choose $N_1(Y)$ so that for all $n > N_1(Y)$ holds: if for some i, j $y_i \in I_j$ then $I_{j-1} \cap Y = I_{j+1} \cap Y = \varnothing$. When for certain i, j $y_i \in I_j$ let $\overline{P}_j(x) := L(x; f; x_{j+1}, y_i, x_{j-2})$, where L is a quadratic Lagrange polynomial interpolating the function f at the points x_{j+1}, y_i, x_{j-2}. For these i, j it satisfies

$$\left\|f - \overline{P}_j\right\|_{O_i} \leqslant c\omega_3(|O_i|), \tag{2.5}$$

and

$$\overline{P}_j(x)\Pi(x) \geqslant 0, \quad x \in O_i. \tag{2.6}$$

Indeed, (2.6) is evident, and (2.5) follows, for example, from [7], Theorem 4.2. Thus Lemma 4 follows from (2.5), (2.6), (1.3) and Lemma 1. □

Now let $I_{i,j}$ be the smallest interval containing I_i and I_j and denote $h_{i,j} := |I_{i,j}|$. Following [6], for $S \in \Sigma_{k-1}$ put

$$a_{i,j} = a_{i,j}(S, \varphi) := \frac{\|p_i - p_j\|_{I_i}}{\varphi(h_j)} \left(\frac{h_j}{h_{i,j}}\right)^k, \qquad i, j = \overline{1, n}, \tag{2.7}$$

where p_i is the polynomial defined by $p_i|_{I_i} := S|_{I_i}$. Finally let

$$a_k := a_k(S, \varphi) := \max_{1 \leqslant i,j \leqslant n} a_{i,j}(S, \varphi),$$

The next Statement 2 which we need was proved by D.Leviatan and I.A.Shevchuk [6] (Lemma 6).

Statement 2. *There is a constant c, depending only on k, such that for any $f \in H_k^\varphi$ and $S \in \Sigma_{k-1}$, if*

$$\|f - S\|_{I_j} \leqslant \varphi(h_j), \qquad j = 1,..., n, \tag{2.8}$$

then

$$a_k \leqslant c. \tag{2.9}$$

We need one statement for the partition of unity, proved in already mentioned paper by J.Gilewicz and I.A.Shevchuk [2].

Statement 3. *For each fixed integer l there exists a collection $\{\tau_{j,n}\}_{j=1}^n$ of polynomials of degree $\leqslant cln$ with the following properties:*

$$\sum_{j=1}^{n} \tau_{j,n}(x) \equiv 1, \tag{2.10}$$

$$\left|\tau_{j,n}^{(\lambda)}(x)\right| \leqslant C \frac{h_j}{\rho_n^{\lambda+1}(x)} \left(\frac{\rho_n(x)}{|x - x_j| + \rho_n(x)}\right)^{l+1}, \quad x \in I, \quad \lambda = 0, 1, 2,..., \tag{2.11}$$

where $C = C(s, l, \lambda)$, *and*

$$\tau_{j,n}(y_i) = 0, \qquad \forall i, j \text{ such that } y_i \in \widetilde{O}_q \text{ and } I_j \not\subset \widetilde{O}_q. \tag{2.12}$$

We have to prove one more lemma.

Lemma 5. *Let* $l \geqslant 3k-1$ *and assume that* $S \in \Sigma_{k-1,O}$ *and* $S(y_i) = 0,\ i = 1,..., s$. *Then the polynomial*

$$D_n(x) := \sum_{i=1}^{n} p_i(x)\tau_{i,n}(x), \tag{2.13}$$

satisfies

$$D_n(y_i) = 0, \quad 1 \leqslant i \leqslant s, \tag{2.14}$$

and for each $\lambda = 0,..., s$,

$$\left|S^{(\lambda)}(x) - D_n^{(\lambda)}(x)\right| \leqslant C_0 a_k(S)\frac{\varphi(\rho_n(x))}{\rho_n^{\lambda}(x)}, \quad x \in I, \tag{2.15}$$

where $C_0 = C_0(s, k, l)$.

Remark 1. *By* $S^{(\lambda)}(x_j)$, $j = 1,..., n$ *we mean either* $S^{(\lambda)}(x_j+)$ *or* $S^{(\lambda)}(x_j-)$ *and inequality (2.15) is considered for* $x \to x_j+$ *or* $x \to x_j-$ *respectively. It doesn't matter what derivative we choose. This fact can be seen from the proof of Lemma 8.*

Proof. Evidently (2.14) follows by (2.12), where we use the fact that S is a single polynomial in each connected component of O. Further we fix $1 \leqslant \nu \leqslant n$, and $x \in I_\nu$. Since $p_\nu - p_i$ is a polynomial of degree not exceeding $k-1$, then

$$\|p_\nu - p_i\|_{I_\nu} \leqslant c\left(\frac{h_{i,\nu}}{h_i}\right)^{k-1} \|p_\nu - p_i\|_{I_i}.$$

Hence by (1.7) and (2.7),

$$\begin{aligned} \|p_\nu - p_i\|_{I_\nu} &\leqslant c\left(\frac{h_{i,\nu}}{h_i}\right)^{k-1}\left(\frac{h_{i,\nu}}{h_\nu}\right)^{k}\varphi(h_\nu)a_{i,\nu} \\ &\leqslant ca_{i,\nu}\varphi(h_\nu)\left(\frac{h_{i,\nu}}{h_\nu}\right)^{3k-2} =: c\Omega_{i,\nu} \end{aligned} \tag{2.16}$$

Now by (2.10),

$$S(x) - D_n(x) = \sum_{i\neq\nu}(p_\nu(x) - p_i(x))\tau_i(x) =: \sum_{i\neq\nu}\alpha_i(x), \tag{2.17}$$

and by virtue of (2.11) and (2.16),

$$\begin{aligned} |\alpha_i(x)| &\leqslant c(l,k)a_{i,\nu}h_i\varphi(h_\nu)\left(\frac{h_{i,\nu}}{h_\nu}\right)^{3k-2}\frac{\rho^l}{(\rho + |x - x_i|)^{l+1}} \\ &\leqslant c(l,k)a_k h_i\frac{\varphi(\rho)}{\rho}\left(\frac{\rho}{\rho + |x - x_i|}\right)^{l+3-3k}. \end{aligned}$$

Hence

$$\begin{aligned} |S(x) - D_n(x)| &\leqslant \sum_{i=1, i\neq\nu}^{n} |\alpha_i(x)| \\ &\leqslant c(l,k)a_k\frac{\varphi(\rho)}{\rho}\sum_{i=1}^{n} h_i\left(\frac{\rho}{\rho + |x - x_i|}\right)^{l+3-3k} \end{aligned}$$

Finally by means of (1.6)

$$|S(x) - D_n(x)| \leqslant c(l,k)a_k \frac{\varphi(\rho)}{\rho} \sum_{i=1}^{n} h_i \left(\frac{\rho}{\rho + |x - x_i|} \right)^2 \leqslant c(l,k)a_k \varphi(\rho),$$

what is (2.15) in the case $\lambda = 0$. If $0 < \lambda \leqslant s$ and $x \in I_\nu$, then with the help of representation (see (2.17))

$$S^{(\lambda)}(x) - D^{(\lambda)}(x) = \sum_{i=1, i \neq \nu}^{n} \sum_{q=0}^{\lambda} \binom{\lambda}{q} (p_\nu^{(q)}(x) - p_i^{(q)}(x)) \tau_i^{(\lambda - q)}(x),$$

taking into account (2.11) and observing that it follows by (2.16) that

$$\left\| p_\nu^{(q)} - p_i^{(q)} \right\|_{I_\nu} \leqslant \frac{c}{h_\nu^q} \, \Omega_{i,\nu},$$

we can obtain (2.15) for $\lambda > 0$ by repeating the previous arguments. This concludes the proof. Note that we used no continuity of the spline S in the proof. □

3 Main lemmas

For every $j = \overline{1,n}$ we set

$$x_j^0 := x_{j,n}^0 := \begin{cases} \cos\left((j - 1/4)\pi/n\right), & \text{if } j < n/2, \\ \cos\left((j - 3/4)\pi/n\right), & \text{if } j \geqslant n/2; \end{cases}$$

$$\overline{x}_j := \overline{x}_{j,n} := \cos\left(j - 1/2\right)\pi/n;$$

$$t_j(x) := t_{j,n}(x) := \left(x - x_j^0\right)^{-2} \cos^2 2n \arccos x + \left(x - \overline{x}_j\right)^{-2} \sin^2 2n \arccos x \text{ —} \tag{3.1}$$

the polynomial of degree $4n - 2$; and for $j \in H$, $b \in \mathbf{N}$ denote

$$d_j := d_{j,n}(b, Y) := \min_{x \in I_j} t_j^b(x) \left| \Pi(x) \right|;$$

$$T_j(x) := T_{j,n}(x, b, Y) := \frac{t_j^b(x)\Pi(x)}{d_j} \text{ —}$$

the polynomial of degree $b(4n - 2) + s$. Obviously for $j \in H$ $d_j > 0$, so the definition is correct.

Lemma 6. *Let $j \in H$ and $b \in \mathbf{N}$, $b \geqslant \frac{s+k}{2}$, $\varphi \in \Phi^k$. Then*

$$c h_j^{-2b} \left| \Pi(x_j) \right| \leqslant d_j \leqslant C h_j^{-2b} \left| \Pi(x_j) \right|, \tag{3.2}$$

$$T_j(x)\Pi(x) \geqslant 0, \quad x \in I, \tag{3.3}$$

$$\varphi(h_j) \left| T_j(x) \right| \geqslant \varphi(\rho), \quad x \in I_j, \tag{3.4}$$

$$\varphi(h_j) \left| T_j(x) \right| \geqslant \varphi(\rho) \left(\frac{\rho}{|x - x_j| + \rho} \right)^{4b+k} \left| \frac{\Pi(x)}{\Pi(x_j)} \right|, \quad x \in I, \tag{3.5}$$

$$\varphi(h_j) \left| T_j(x) \right| \leqslant c\varphi(\rho) \left(\frac{\rho}{|x - x_j| + \rho} \right)^{\frac{2b-s-k}{2}}, \quad x \in I. \tag{3.6}$$

Proof. For d_j we have by (1.8) (see also [7], Proof of Lemma 17.2)

$$t_j^b(x)\,|\Pi(x)| \geqslant c\min\{(x-x_j^0)^{-2b},(x-\overline{x}_j)^{-2b}\}\,|\Pi(x_j)| \geqslant ch_j^{-2b}\,|\Pi(x_j)|\,,\quad x\in I_j.$$

It is known that

$$t_j(x) \leqslant 312h_j^{-2},\quad x\in I_j.$$

So the above estimate of d_j is

$$t_j^b(x)\,|\Pi(x)| \leqslant Ch_j^{-2b}\,|\Pi(x_j)|\,,\quad x\in I_j.$$

Also we have

$$\max\left\{\frac{1}{(x-x_j^0)^2},\frac{1}{(x-\overline{x}_j)^2}\right\} \leqslant c\left(\frac{1}{|x-x_j|+h_j}\right)^2,\quad x\notin I_j, \tag{3.7}$$

and

$$\min\left\{\frac{1}{(x-x_j^0)^2},\frac{1}{(x-\overline{x}_j)^2}\right\} \geqslant c\left(\frac{1}{|x-x_j|+h_j}\right)^2,\quad x\notin I_j. \tag{3.8}$$

And for the polynomial T_j we obtain

$$\varphi(h_j)\,|T_j(x)| \geqslant c\varphi(\rho),\quad x\in I_j;$$

taking into account the estimates of d_j, (3.7) and (1.2)–(1.8), for $x\in I$

$$\begin{aligned}
\varphi(h_j)\,|T_j(x)| &\leqslant c\varphi(h_j)\left(\frac{h_j}{|x-x_j|+h_j}\right)^{2b}\left|\frac{\Pi(x)}{\Pi(x_j)}\right| \\
&\leqslant c\varphi(h_j)\left(\frac{h_j}{|x-x_j|+h_j}\right)^{2b-s} \\
&\leqslant c\varphi\left(\sqrt{4\rho(|x-x_j|+\rho)}\right)\left(\frac{h_j}{|x-x_j|+h_j}\right)^{2b-s} \\
&\leqslant c\varphi(\rho)\left(\frac{|x-x_j|+\rho}{\rho}\right)^{k/2}\left(\frac{h_j}{|x-x_j|+h_j}\right)^{2b-s} \\
&\leqslant c\varphi(\rho)\left(\frac{h_j}{|x-x_j|+h_j}\right)^{2b-s-k} \\
&\leqslant c\varphi(\rho)\left(\frac{\rho}{|x-x_j|+\rho}\right)^{\frac{2b-s-k}{2}}.
\end{aligned}$$

Applying (3.2), (3.8) and (1.2)–(1.8), for $x\notin I_j$

$$\begin{aligned}
\varphi(h_j)\,|T_j(x)| &\geqslant c\varphi(h_j)\left(\frac{h_j}{|x-x_j|+h_j}\right)^{2b}\left|\frac{\Pi(x)}{\Pi(x_j)}\right| \\
&\geqslant c\varphi\left(\frac{\rho^2}{|x-x_j|+h_j}\right)\left(\frac{h_j}{|x-x_j|+h_j}\right)^{2b}\left|\frac{\Pi(x)}{\Pi(x_j)}\right| \\
&\geqslant c\varphi\left(\frac{\rho^2}{|x-x_j|+\rho}\right)\left(\frac{h_j}{|x-x_j|+h_j}\right)^{2b}\left|\frac{\Pi(x)}{\Pi(x_j)}\right| \\
&\geqslant c\varphi(\rho)\left(\frac{\rho}{|x-x_j|+\rho}\right)^{k}\left(\frac{h_j}{|x-x_j|+h_j}\right)^{2b}\left|\frac{\Pi(x)}{\Pi(x_j)}\right| \\
&\geqslant c\varphi(\rho)\left(\frac{\rho}{|x-x_j|+\rho}\right)^{4b+k}\left|\frac{\Pi(x)}{\Pi(x_j)}\right|.
\end{aligned}$$

Now the proof is complete. □

In the next lemma we construct correcting polynomials.

Lemma 7. *For each $\varphi \in \Phi^k$, there exists a polynomial $Q_n = Q_n(x, \varphi)$ of degree $\leqslant (2s + 2k + 16)n$, satisfying*

$$Q_n(x)\Pi(x) \geqslant 0, \qquad x \in I; \tag{3.9}$$

$$Q_n(x)\operatorname{sgn}\Pi(x) \geqslant \varphi(\rho), \qquad x \in I \setminus O; \tag{3.10}$$

$$|Q_n(x)| \leqslant c_1\varphi(\rho), \qquad x \in I. \tag{3.11}$$

Proof. We take a polynomial $T_j(x) = T_j(x, b, Y)$, defined above the Lemma 6, where $b = \left[\frac{s+k}{2}\right] + 4$. Put

$$Q_n(x) := \sum_{j\in H} \varphi(h_j)T_j(x).$$

Now (3.9) follows from (3.3). Because at the fixed x all $T_j(x)$ have the same sign, (3.10) follows from (3.4). We have also by (3.6)

$$\varphi(h_j)\,|T_j(x)| \leqslant c\varphi(\rho)\left(\frac{\rho}{|x - x_j| + \rho}\right)^3.$$

Hence we obtain at last with the help of (1.2), (1.5) and (1.6)

$$|Q_n(x)| \leqslant c\sum_{j=1}^{n} \varphi(\rho)\left(\frac{\rho}{|x - x_j| + \rho}\right)^3$$

$$\leqslant c\sum_{j=1}^{n} \varphi(\rho)h_j\rho^{-1}\left(\frac{\rho}{|x - x_j| + \rho}\right)^2 \leqslant c_1\varphi(\rho),$$

that is (3.11). □

4 Proofs of the positive results

Lemma 8. *Let $\varphi \in \Phi^k$ and $S \in \Sigma_{k-1,O}$. Assume that*

$$a_k(S, \varphi) \leqslant c_2, \tag{4.1}$$

and

$$S(x)\Pi(x) \geqslant 0, \qquad x \in I \tag{4.2}$$

Then there is a polynomial P_n of degree $\leqslant cn$ such that

$$|P_n(x) - S(x)| \leqslant C\varphi(\rho_n(x)), \tag{4.3}$$

and

$$P_n(x)\Pi(x) \geqslant 0, \qquad x \in I. \tag{4.4}$$

Proof. We begin with the polynomial D_n of Lemma 5, associated with the spline S (we take $l = 3k - 1$) of degree $\leqslant cn$. We have

$$|S(x) - D_n(x)| \leqslant C_0c_2\varphi(\rho) \tag{4.5}$$

Now we proceed with the polynomial of degree $\leqslant cn$

$$R_n := D_n + C_0c_2Q_n, \tag{4.6}$$

where the polynomial Q_n is defined in Lemma 7. By virtue of (3.11) and (4.5) we have

$$|S(x) - R_n(x)| \leqslant c\varphi(\rho), \qquad x \in I,$$

and by virtue of (3.10), (4.2) and (4.5)

$$\begin{aligned} R_n(x)\operatorname{sgn}\Pi(x) \geqslant C_0c_2Q_n(x)\operatorname{sgn}\Pi(x) + S(x)\operatorname{sgn}\Pi(x) \\ - |S(x) - D_n(x)| \geqslant 0, \quad x \in I \setminus O. \end{aligned}$$

It remains to modify the polynomial R_n so that it will be copositive with f also in O. For this we have deal only with D_n, because by (3.9) copositivity takes place for Q_n. So we take $\widetilde{O}_q$, a connected component of closure $\overline{O}$ of O and we let $y_{i_q+1},\dots, y_{i_q+\nu} \in \widetilde{O}_q$ ($\nu \leqslant s$). Also we put x_{j_q} to be the closest to $\widetilde{O}_q$, such that $I_{j_q} \cap O = \varnothing$. Then (2.14) and (2.15) together with (4.1) and (4.2) yield for $x \in \widetilde{O}_q$

$$\begin{aligned} |S(x) - D_n(x)| &= \left|\prod_{\mu=1}^{\nu}(x - y_{i_q+\mu})[x, y_{i_q+1},\dots, y_{i_q+\nu}; S - D_n]\right| \\ &= \frac{1}{(\nu+1)!}\left|\prod_{\mu=1}^{\nu}(x - y_{i_q+\mu})(S^{(\nu)}(\theta) - D_n^{(\nu)}(\theta))\right| \\ &\leqslant c\varphi(\rho_n(\theta))\rho_n^{-\nu}(\theta)\left|\prod_{\mu=1}^{\nu}(x - y_{i_q+\mu})\right|, \end{aligned} \tag{4.7}$$

by the square brackets we denoted the divided difference of order ν of the function $S - D_n$ and $\theta \in \widetilde{O}_q$. For the definition and necessary properties of divided differences one can see [7], ch.1. Hence

$$\begin{aligned} |S(x) - D_n(x)| &\leqslant c\varphi(\rho_n(x_{j_q}))\frac{\prod_{\mu=1}^{\nu}|x - y_{i_q+\mu}|}{\prod_{\mu=1}^{\nu}|x_{j_q} - y_{i_q+\mu}|} \\ &\leqslant c_3\varphi(\rho_n(x_{j_q}))\frac{|\Pi(x, Y)|}{|\Pi(x_{j_q}, Y)|}. \end{aligned} \tag{4.8}$$

For the first inequality in (4.8) we used the fact that $\widetilde{O}_q$ is connected and contains at most $3s$ intervals, so that

$$\left|\widetilde{O}_q\right| \sim \rho_n(\theta) \sim \rho_n(x_{j_q}) \sim (x_{j_q} - y_{j_q+\mu}), \quad 1 \leqslant \mu \leqslant \nu;$$

and for the second inequality we used the above together with the fact that for any $y_i \notin \widetilde{O}_q$

$$\frac{|x - y_i|}{|x_{j_q} - y_i|} \geqslant \frac{|x - y_i|}{|x - x_{j_q}| + |x - y_i|} \geqslant \frac{\rho_n(x)}{|\widetilde{O}_q| + \rho_n(x)} \geqslant c > 0.$$

Now we again make use of polynomials T_j from Lemma 6 (and again we take $b = \left[\frac{s+k}{2}\right]+4$). By (3.5) we have

$$\varphi(h_{j_q})\left|T_{j_q}(x)\right| \geqslant c_4\varphi(\rho_n(x_{j_q}))\frac{|\Pi(x, Y)|}{|\Pi(x_{j_q}, Y)|}, \quad x \in \widetilde{O}_q. \tag{4.9}$$

Thus if we set

$$\widetilde{T}_{j_q} := \frac{c_3}{c_4}\varphi(h_{j_q})T_{j_q},$$

and

$$\widetilde{D}_n := D_n + \sum_q \widetilde{T}_{j_q},$$

where the sum is taken over all connected components $\widetilde{O}_q$, then it follows by (4.8), (4.9), (3.3) and (4.2), that for $x \in \widetilde{O}_{q_0}$

$$\widetilde{D}_n(x)\Pi(x,Y) = (D_n(x) - S(x))\Pi(x,Y) + \widetilde{T}_{j_{q_0}}(x)\Pi(x,Y) + \sum_{q \neq q_0} \widetilde{T}_{j_q}(x)\Pi(x,Y) + S(x)\Pi(x,Y) \geqslant 0.$$

Hence, for the polynomial

$$P_n := R_n + \sum_q \widetilde{T}_{j_q},$$

we have

$$P_n(x)\Pi(x,Y) \geqslant 0, \qquad x \in I.$$

Note that one can prove that

$$\left| \sum_q \widetilde{T}_{j_q}(x) \right| \leqslant c_5 \varphi(\rho_n(x)), \quad x \in I$$

as in the proof of Lemma 7. So we observe that

$$|S(x) - P_n(x)| \leqslant c_6 \varphi(\rho_n(x)), \quad x \in I.$$

This concludes the proof. □

Proofs of Theorems 1 and 2. Theorem 1 for $n \geqslant c_7 \geqslant k + r - 1$ now follows from Lemmas 2 and 8 with the help of Statement 2. While for $k+r-1 \leqslant n < c_7$ it follows from the Statement 1. Theorem 2 for $n \geqslant N(Y)$ follows from Lemmas 3, 4, 8 and Statement 2, and for $k + r - 1 \leqslant n < N(Y)$ it follows also from Statement 1. □

5 Counterexamples

Example 1. *For each $s > 1$, $r < s - 1$, $A > 0$, $n \in \mathbf{N}$ there exists a collection $Y := Y(n, r, A, s)$ and function $f := f_{n,r,A} \in C^{(r)} \cap \Delta^{(0)}(Y)$ such that*

$$E_n^{(0)}(f, Y) \geqslant A\omega_2(f^{(r)}; 1) \geqslant A2^{-k+2}\omega_k(f^{(r)}; 1), \quad k \geqslant 2. \tag{5.1}$$

Proof. We assume without loss of generality that $n \geqslant r + 1$. Let choose $b \in (0, 1)$ from the condition

$$\frac{1}{4bn^{2(r+1)}} - \frac{b^r}{4(r+1)!} = A.$$

We fix arbitrary collection Y of points y_i such that $-1 + b = y_1 > y_2 > ... > y_s > -1$. Following [2], set

$$Q_{r+1}(x) := (x - y_1)^{r+1};$$

$$f(x) := (x - y_1)_+^{r+1} := \begin{cases} Q_{r+1}(x), & \text{if } x \geqslant -1 + b, \\ 0, & \text{if } x < -1 + b. \end{cases}$$

Evidently $f \in C^{(r)} \cap \Delta^{(0)}(Y)$. We take an arbitrary polynomial $P_n \in \mathcal{P}_n \cap \Delta^{(0)}(Y)$. Let's denote

$$R_n(x) := Q_{r+1}(x) - P_n(x).$$

We consider divided difference

$$[y_1,..., y_{r+2}; R_n].$$

Because $P_n \in \Delta^{(0)}(Y)$, there is $P_n(y_i) = 0,\ i = \overline{1, r+2}$, so $[y_1,..., y_{r+2}; P_n] = 0$. Besides, $[y_1,..., y_{r+2}; Q_{r+1}] = 1$, that is,

$$[y_1,..., y_{r+2}; R_n] = 1.$$

Therefore the point $\theta \in (-1, -1 + b)$ exists such that

$$R_n^{(r+1)}(\theta) = (r+1)![y_1,..., y_{r+2}; R_n] = (r+1)!.$$

Applying Markov's inequality we obtain

$$\begin{aligned}(r+1)! = R_n^{(r+1)}(\theta) \leqslant n^{2(r+1)}\,\|R_n\| &\leqslant n^{2(r+1)}(\|f - P_n\| + \|f - Q_{r+1}\|) \\ &= n^{2(r+1)}(\|f - P_n\| + b^{r+1}),\end{aligned}$$

whence

$$\|f - P_n\| > \frac{(r+1)!}{n^{2(r+1)}} - b^{r+1}.$$

On the other hand ,

$$\omega_2(f^{(r)}; 1) = \omega_2(f^{(r)} - Q_{r+1}^{(r)}; 1) \leqslant 4\left\|f^{(r)} - Q_{r+1}^{(r)}\right\| = 4(r+1)!b.$$

Hence

$$\frac{\|f - P_n\|}{\omega_2(f^{(r)}; 1)} > \frac{1}{4bn^{2(r+1)}} - \frac{b^r}{4(r+1)!} = A.$$

□

Example 2. *Let $r = s - 1$. For each $A > 0$, $n \in \mathbf{N}$ there exists a collection $Y := Y(n, r, A)$ and function $f := f_{n,r,A} \in C^{(r)} \cap \Delta^{(0)}(Y)$ such that*

$$E_n^{(0)}(f, Y) \geqslant A\omega_3(f^{(r)}; 1) \geqslant A2^{-k+3}\omega_k(f^{(r)}; 1), \quad k \geqslant 3. \tag{5.2}$$

Proof. We assume that $n \geqslant r + 2$. Let choose $b \in (0, 1)$ from the condition

$$\frac{1}{4(r+2)bn^{2(r+1)}} - \frac{b^r}{4(r+2)!} = A.$$

We fix arbitrary collection Y of points y_i such that $-1 + b = y_1 > y_2 > ... > y_s > -1$. Set

$$Q_{r+2}(x) := (x - y_1)^{r+2};$$

$$f(x) := (x - y_1)_+^{r+2} := \begin{cases} Q_{r+2}(x), & \text{if } x \geqslant -1 + b, \\ 0, & \text{if } x < -1 + b. \end{cases}$$

Evidently $f \in C^{(r)} \cap \Delta^{(0)}(Y)$. We take an arbitrary polynomial $P_n \in \mathcal{P}_n \cap \Delta^{(0)}(Y)$. Let's denote

$$R_n(x) := Q_{r+2}(x) - P_n(x).$$

We set $y_{r+2} := -1$ and consider divided difference

$$[y_1,...,y_{r+2}; R_n].$$

Because $P_n \in \Delta^{(0)}(Y)$, there is $P_n(y_i) = 0,\ i = \overline{1, r+1}$, so

$$[y_1,...,y_{r+2}; P_n] = \frac{P_n(-1)}{\Pi(-1)} \geqslant 0.$$

Besides, $[y_1,...,y_{r+2}; Q_{r+2}] \leqslant 0$, and (see [7], Theorem 1.2)

$$-[y_1,...,y_{r+2}; Q_{r+2}] \geqslant (y_1 - y_2) + ... + (y_{r+1} - y_{r+2}) = b$$

that is,

$$[y_1,...,y_{r+2}; R_n] \leqslant -b.$$

Therefore the point $\theta \in (-1, -1 + b)$ exists such that

$$\left|R_n^{(r+1)}(\theta)\right| = (r+1)!\,|[y_1,...,y_{r+2}; R_n]| \geqslant (r+1)!b.$$

Applying Markov's inequality we obtain

$$\begin{aligned}(r+1)!b \leqslant \left|R_n^{(r+1)}(\theta)\right| \leqslant n^{2(r+1)}\,\|R_n\| &\leqslant n^{2(r+1)}(\|f - P_n\| + \|f - Q_{r+2}\|) \\ &= n^{2(r+1)}(\|f - P_n\| + b^{r+2}),\end{aligned}$$

whence

$$\|f - P_n\| > \frac{(r+1)!b}{n^{2(r+1)}} - b^{r+2}.$$

On the other hand ,

$$\omega_3(f^{(r)}; 1) = \omega_3(f^{(r)} - Q_{r+2}^{(r)}; 1) \leqslant 8\left\|f^{(r)} - Q_{r+2}^{(r)}\right\| = 4(r+2)!b^2.$$

Hence

$$\frac{\|f - P_n\|}{\omega_3(f^{(r)}; 1)} > \frac{1}{4(r+2)bn^{2(r+1)}} - \frac{b^r}{4(r+2)!} = A.$$

□

Example 3. *For any collection $Y \in \mathbf{Y}_s$ there is a function $f = f(Y) \in \Delta^{(0)}(Y)$ such that*

$$\limsup_{n\to\infty} \frac{E_n^{(0)}(f, Y)}{\omega_k(f, 1/n)} = \infty, \tag{5.3}$$

when $k \geqslant 4$.

Proof. Note that it is sufficient to prove (5.3) only for $k = 4$, so we will deal with the fourth modulus. Denote

$$\Pi_1(x) := \begin{cases} \prod_{i=2}^{s}(x - y_i), & s \geqslant 2, \\ 1, & s = 1. \end{cases}$$

For $b \in [0, 1]$ we construct a function S_b such that

$$S_b(x) = \begin{cases} 1, & |x| \geqslant 2b; \\ 0, & |x| \leqslant b; \end{cases}$$

$$0 \leqslant S_b(x) \leqslant 1, \quad b < |x| < 2b; \qquad S_b \in C^{(4)}.$$

For example for $x \in [b, 2b]$ one can set

$$S_b(x) = \int_b^x (u-b)^4(2b-u)^4 du \left(\int_b^{2b} (u-b)^4(2b-u)^4 du \right)^{-1},$$

and $S_b(x) = S_b(-x),\ x \in [-2b, -b]$. Put

$$q_b(x) := ((x-y_1)^2 - b^2)(x-y_1),$$
$$g_b(x) := q_b(x) S_b(x-y_1) \frac{\Pi_1(x)}{\|\Pi_1\|}, \qquad Q_b(x) := q_b(x) \frac{\Pi_1(x)}{\|\Pi_1\|}.$$

For these functions there are following relations:

$$g_b(x) = 0, \quad x \in [y_1 - b, y_1 + b], \tag{5.4}$$
$$g_b \in C^{(4)} \cap \Delta^{(0)}(Y), \qquad \|g_b\| \leqslant \|Q_b\| \leqslant \|q_b\| \leqslant 10, \tag{5.5}$$
$$\|g_b - Q_b\|_{[y_1, y_1+2b]} \leqslant 6b^3, \tag{5.6}$$

evidently $0 < B = \frac{\Pi_1(y_1)}{\|\Pi_1\|} < 1$, therefore

$$\omega_4(g_b; t) \leqslant 16 \|g_b - Q_b\| + t^4 \left\| Q_b^{(4)} \right\| \leqslant 96b^3 + 10t^4(s+2)^8 \leqslant c_7(b^3 + t^4). \tag{5.7}$$

For each $P_n \in \mathcal{P}_n \cap \Delta^{(0)}(Y)$ we denote

$$R_n(x) = P_n(x) - Q_b(x),$$

and obtain

$$P_n'(y_1) \geqslant 0, \quad R_n'(y_1) \geqslant -Q_b'(y_1) = Bb^2. \tag{5.8}$$

Applying Bernstein's inequality (we assume $n > s+1$), for $0 < h \leqslant 1 - |y_1|$ from (5.6) and (5.8)

$$Bhb^2 \leqslant hR_n'(y_1) \leqslant n \|R_n\|_{[y_1-h, y_1+h]} \leqslant n(\|g_b - P_n\|_{[y_1-h, y_1+h]} + 6b^3).$$

Hence

$$\|g_b - P_n\|_{[y_1-h, y_1+h]} \geqslant B\frac{hb^2}{n} - 6b^3. \tag{5.9}$$

For $n > s+1$ put

$$b_n = n^{-4/3}; \quad f_n(x) = g_{b_n}(x) \tag{5.10}$$

and note that (5.7) implies

$$\omega_4(f_n; t) \leqslant 2c_7 t^4,\ t \geqslant 1/n. \tag{5.11}$$

First we put $\varepsilon = 0.1$ and choose n_0 so large that $n_0^{\varepsilon} > 4c_7$ and $b_{n_0} \leqslant \frac{1-|y_1|}{2}$. Set $d_0 := 1$ and

$$d_j := \frac{B}{40} \frac{b_{n_{j-1}} b_{n_j}^2}{n_j} d_{j-1}, \tag{5.12}$$

where the increasing sequence $\{n_\nu\}_{\nu=1}^{\infty}$ is defined by induction as follows. Suppose that $\{n_0, ..., n_{\sigma-1}\}$ have been defined, then put

$$F_{\sigma-1} := \sum_{j=1}^{\sigma-1} d_{j-1} f_{n_j} \ (F_0 := 0)$$

and take $n_\sigma > n_{\sigma-1}$ so large that

$$\left\|F^{(4)}_{\sigma-1}\right\| < d_{\sigma-1}n^{\varepsilon}_{\sigma}, \tag{5.13}$$

and

$$Bb_{n_{\sigma-1}} > 12n^{-\varepsilon}_{\sigma}. \tag{5.14}$$

We will put

$$\Phi_\sigma = \sum_{j=\sigma}^{\infty} d_{j-1}f_{n_j}, \tag{5.15}$$

the uniform convergence of the series follows from (5.5) and the inequality

$$\sum_{j=\sigma}^{\infty} d_{j-1} < d_{\sigma-1}\left(1 + \frac{1}{40} + \frac{1}{40^2} + ...\right) < 2d_{\sigma-1}, \tag{5.16}$$

which follows from (5.12), (5.10) and $0 < B < 1$. Next we define

$$f := f(Y) := \sum_{j=1}^{\infty} d_{j-1}f_{n_j}, \tag{5.17}$$

and note that (5.5) yields that $f \in \Delta^{(0)}(Y)$. Inequalities (5.11), (5.16) and $n^{\varepsilon}_0 > 4c_7$ yield

$$\omega_4\left(\Phi_\sigma; \frac{1}{n_\sigma}\right) \leqslant 2c_7\frac{1}{n^4_\sigma}\sum_{j=\sigma}^{\infty} d_{j-1} < 4c_7 d_{\sigma-1}\frac{1}{n^4_\sigma} < d_{\sigma-1}\frac{1}{n^{4-\varepsilon}_\sigma},$$

and (5.13) provides

$$\omega_4\left(F_{\sigma-1}; \frac{1}{n_\sigma}\right) \leqslant \frac{1}{n^4_\sigma}\left\|F^{(4)}_{\sigma-1}\right\| < d_{\sigma-1}\frac{1}{n^{4-\varepsilon}_\sigma}.$$

So for all σ

$$\omega_4\left(f; \frac{1}{n_\sigma}\right) < \frac{2d_{\sigma-1}}{n^{4-\varepsilon}_\sigma}. \tag{5.18}$$

By (5.4) $F_{\sigma-1}$ is zero on $[y_1 - b_{n_{\sigma-1}}, y_1 + b_{n_{\sigma-1}}] =: J_\sigma$, so we may write

$$f(x) = d_{\sigma-1}f_{n_\sigma}(x) + \Phi_{\sigma+1}(x), \quad x \in J_\sigma.$$

For each $P_n \in \mathcal{P}_n \cap \Delta^{(0)}(Y)$ let $p_{n_\sigma} = 1/d_{\sigma-1}P_{n_\sigma}$. Since $p_{n_\sigma} \in \Delta^{(0)}(Y)$, then by means of (5.9)

$$\|p_{n_\sigma} - f_{n_\sigma}\|_{J_\sigma} \geqslant B\frac{b_{n_{\sigma-1}}b^2_{n_\sigma}}{n_\sigma} - 6b^3_{n_\sigma}.$$

On the other hand (5.5), (5.16), and (5.12) yield

$$\|\Phi_{\sigma+1}\| \leqslant 10\sum_{j=\sigma+1}^{\infty} d_{j-1} \leqslant 20d_\sigma = \frac{1}{2}d_{\sigma-1}B\frac{b_{n_{\sigma-1}}b^2_{n_\sigma}}{n_\sigma}.$$

Hence

$$\begin{aligned}\|f - P_{n_\sigma}\| &\geqslant \|f - P_{n_\sigma}\|_{J_\sigma} \geqslant \|P_{n_\sigma} - d_{\sigma-1}f_{n_\sigma}\|_{J_\sigma} - \|\Phi_{\sigma+1}\| \\ &= d_{\sigma-1}\|p_{n_\sigma} - f_{n_\sigma}\|_{J_\sigma} - \|\Phi_{\sigma+1}\| \\ &\geqslant d_{\sigma-1}\left(\frac{1}{2}B\frac{b_{n_{\sigma-1}}b^2_{n_\sigma}}{n_\sigma} - 6b^3_{n_\sigma}\right).\end{aligned}$$

Now by (5.10) and (5.14)

$$\|f - P_{n_\sigma}\| \geqslant 6d_{\sigma-1}\left(\frac{b_{n_\sigma}^2}{n_\sigma^{1+\varepsilon}} - b_{n_\sigma}^3\right) = 6d_{\sigma-1}\left(\left(\frac{1}{n_\sigma}\right)^{11/3+\varepsilon} - \left(\frac{1}{n_\sigma}\right)^4\right).$$

And by (5.18)

$$\frac{\|f - P_{n_\sigma}\|}{\omega_4(f; 1/n_\sigma)} \geqslant 3\big(n_\sigma^{1/3-2\varepsilon} - n_\sigma^{-\varepsilon}\big) \to \infty, \quad \sigma \to \infty,$$

which implies (5.3). □

Acknowledgment

In conclusion I express a sincere gratitude to Professor I.A.Shevchuk for the statement of the problem, constant attention, useful advice and discussions of the results received in the work and to A.Bondarenko for the expressed remarks.

References

[1] G.A.Dzyubenko, Copositive pointwise approximation (in Russian), *Ukainian Math. Journal*, **48**, 326-334 (1996).

[2] J.Gilewicz and I.A.Shevchuk, Comonotone approximation (in Russian), *Fundamentalnaya i Prikladnaya Mathematica*, **2**, 319-363 (1996).

[3] Y.K.Hu, K.Kopotun, X.M.Yu, Constrained approximation in Sobolev spaces, *Canadian Journal Math.*, **49**, 74-99 (1997).

[4] Y.K.Hu and X.M.Yu, Copositive polynomial approximation revisited, in *Applied Mathematics Reviews* (G.Anastassiou, ed.), World Scientific Pub., 2000, Vol.1, pp.157-179.

[5] D.Leviatan, The degree of copositive approximation by polynomials, *Proc. Amer. Math. Soc.*, **88**, 101-105 (1983).

[6] D.Leviatan and I.A.Shevchuk, Nearly comonotone approximation, *Journal of Approximation Theory*, **95**, 53-81 (1998).

[7] I.A.Shevchuk, *"Polynomial Approximation and Traces of Functions Continuous on a Segment"* (in Russian), Naukova Dumka, Kyiv, 1992.

[8] S.P.Zhou, A counterexample in copositive approximation, *Israel Journal Math.*, **78**, 75-83 (1992).

Instructions to Contributors
Journal of Concrete and Applicable Mathematics

A quartely international publication of Eudoxus Press, LLC, of TN.

Editor in Chief: George Anastassiou
Department of Mathematical Sciences
University of Memphis
Memphis, TN 38152-3240, U.S.A.

1. Manuscripts hard copies in triplicate, and in English, should be submitted to the Editor-in-Chief:

Prof.George A. Anastassiou
Department of Mathematical Sciences
The University of Memphis
Memphis,TN 38152, USA.
Tel.001. 901.678.3144
e-mail:ganastss@memphis.edu.

Authors may want to recommend an associate editor the most related to the submission to possibly handle it.
Also authors may want to submit a list of six possible referees, to be used in case we cannot find related referees by ourselves.
2. Manuscripts should be typed using any of TEX,LaTEX,AMS-TEX,or AMS-LaTEX and according to EUDOXUS PRESS, LLC. LATEX STYLE FILE. (VISIT www.msci.memphis.edu/~ganastss/jcaam / to save a copy of the style file.)They should be carefully prepared in all respects. Submitted copies should be brightly printed (not dot-matrix), double spaced, in ten point type size, on one side high quality paper 8(1/2)x11 inch. Manuscripts should have generous margins on all sides and should not exceed 24 pages.
3. Submission is a representation that the manuscript has not been published previously in this or any other similar form and is not currently under consideration for publication elsewhere. A statement transferring from the authors(or their employers,if they hold the copyright) to Eudoxus Press, LLC, will be required before the manuscript can be accepted for publication.The Editor-in-Chief will supply the necessary forms for this transfer.Such a written transfer of copyright,which previously was assumed to be implicit in the act of submitting a manuscript,is necessary under the U.S.Copyright Law in order for the publisher to carry through the dissemination of research results and reviews as widely and effective as possible.
4. The paper starts with the title of the article, author's name(s) (no titles or degrees), author's affiliation(s) and e-mail addresses. The affiliation should comprise the department, institution (usually university or company), city, state (and/or nation) and mail code.
The following items, 5 and 6, should be on page no. 1 of the paper.

5. An abstract is to be provided, preferably no longer than 150 words.
6. A list of 5 key words is to be provided directly below the abstract. Key words should express the precise content of the manuscript, as they are used for indexing purposes. The main body of the paper should begin on page no. 1, if possible.
7. All sections should be numbered with Arabic numerals (such as: 1. INTRODUCTION) .
Subsections should be identified with section and subsection numbers (such as 6.1. Second-Value Subheading).
If applicable, an independent single-number system (one for each category) should be used to label all theorems, lemmas, propositions, corrolaries, definitions, remarks, examples, etc. The label (such as Lemma 7) should be typed with paragraph indentation, followed by a period and the lemma itself.
8. Mathematical notation must be typeset. Equations should be numbered consecutively with Arabic numerals in parentheses placed flush right,and should be thusly referred to in the text [such as Eqs.(2) and (5)]. The running title must be placed at the top of even numbered pages and the first author's name, et al., must be placed at the top of the odd numbed pages.
9. Illustrations (photographs, drawings, diagrams, and charts) are to be numbered in one consecutive series of Arabic numerals. The captions for illustrations should be typed double space. All illustrations, charts, tables, etc., must be embedded in the body of the manuscript in proper, final, print position. In particular, manuscript, source, and PDF file version must be at camera ready stage for publication or they cannot be considered.
Tables are to be numbered (with Roman numerals) and referred to by number in the text. Center the title above the table, and type explanatory footnotes (indicated by superscript lowercase letters) below the table.
10. List references alphabetically at the end of the paper and number them consecutively. Each must be cited in the text by the appropriate Arabic numeral in square brackets on the baseline.

References should include (in the following order):
initials of first and middle name, last name of author(s)
title of article,
name of publication, volume number, inclusive pages, and year of publication.
Authors should follow these examples:

Journal Article

1. H.H.Gonska,Degree of simultaneous approximation of bivariate functions by Gordon operators, (journal name in italics) *J. Approx. Theory*, 62,170-191(1990).

Book

2. G.G.Lorentz, (title of book in italics) *Bernstein Polynomials* (2nd ed.), Chelsea,New York,1986.

Contribution to a Book

3. M.K.Khan, Approximation properties of beta operators,in(title of book in italics) *Progress in Approximation Theory* (P.Nevai and A.Pinkus,eds.), Academic Press, New York,1991,pp.483-495.

11. All acknowledgements (including those for a grant and financial support) should occur in one paragraph that directly precedes the References section.
12. Footnotes should be avoided. When their use is absolutely necessary, footnotes should be numbered consecutively using Arabic numerals and should be typed at the bottom of the page to which they refer. Place a line above the footnote, so that it is set

off from the text. Use the appropriate superscript numeral for citation in the text.

13. After each revision is made please again submit three hard copies of the revised manuscript, including in the final one. And after a manuscript has been accepted for publication and with all revisions incorporated, manuscripts, including the TEX/LaTex source file and the PDF file, are to be submitted to the Editor's Office on a personal-computer disk, 3.5 inch size. Label the disk with clearly written identifying information and properly ship, such as:

Your name, title of article, kind of computer used, kind of software and version number, disk format and files names of article, as well as abbreviated journal name.

Package the disk in a disk mailer or protective cardboard. Make sure contents of disks are identical with the ones of final hard copies submitted!

Note: The Editor's Office cannot accept the disk without the accompanying matching hard copies of manuscript. No e-mail final submissions are allowed! The disk submission must be used.

14. Effective 1 Nov. 2005 the journal's page charges are $8.00 per PDF file page. Upon acceptance of the paper an invoice will be sent to the contact author. The fee payment will be due one month from the invoice date. The article will proceed to publication only after the fee is paid. The charges are to be sent, by money order or certified check, in US dollars, payable to Eudoxus Press, LLC, to the address shown in the Scope and Prices section.

No galleys will be sent and the contact author will receive one(1) electronic copy of the journal issue in which the article appears.

15. This journal will consider for publication only papers that contain proofs for their listed results.

TABLE OF CONTENTS,JOURNAL OF CONCRETE AND APPLICABLE MATHEMATICS,VOL.3,NO.2,2005

ON A NEW FAMILY OF GAUSSIAN QUADRATURE FORMULAE OF BIRKHOFF TYPE WITH APPLICATIONS TO POLYNOMIAL INEQUALITIES, A.GUESSAB,……..127

A CYCLIC SUBGRAPH METHODOLOGY FOR ESTIMATING DE BRUIJN WEIGHT CLASS DISTRIBUTIONS,G.MAYHEW,…….169

A POSITIVE LINEAR OPERATOR FOR THE APPROXIMATION OF FUNCTIONS OF TWO VARIABLES,A.CIUPA,…….187

ON THE SOLUTION STRUCTURES OF A CLASS OF SEMILINEAR ELLIPTIC EQUATIONS ON R^d,N.RHOUMA,…….199

ON THE DEGREE OF COPOSITIVE APPROXIMATION,I.SMAZHENKO,……221

VOLUME 3,NUMBER 3 JULY 2005

ISSN:1548-5390 PRINT,1559-176X ONLINE



JOURNAL

OF CONCRETE

AND APPLICABLE

MATHEMATICS

EUDOXUS PRESS,LLC

SCOPE AND PRICES OF THE JOURNAL
Journal of Concrete and Applicable Mathematics

A quartely international publication of **Eudoxus Press,LLC**

Editor in Chief: George Anastassiou
Department of Mathematical Sciences,
University of Memphis
Memphis, TN 38152, U.S.A.
ganastss@memphis.edu

The main purpose of the "Journal of Concrete and Applicable Mathematics" is to publish high quality original research articles from all subareas of Non-Pure and/or Applicable Mathematics and its many real life applications, as well connections to other areas of Mathematical Sciences, as long as they are presented in a Concrete way. It welcomes also related research survey articles and book reviews.A sample list of connected mathematical areas with this publication includes and is not restricted to: Applied Analysis, Applied Functional Analysis, Probability theory, Stochastic Processes, Approximation Theory, O.D.E, P.D.E, Wavelet, Neural Networks,Difference Equations, Summability, Fractals, Special Functions, Splines, Asymptotic Analysis, Fractional Analysis, Inequalities, Moment Theory, Numerical Functional Analysis,Tomography, Asymptotic Expansions, Fourier Analysis, Applied Harmonic Analysis, Integral Equations, Signal Analysis, Numerical Analysis, Optimization, Operations Research, Linear Programming, Fuzzyness, Mathematical Finance, Stochastic Analysis, Game Theory, Math.Physics aspects, Applied Real and Complex Analysis, Computational Number Theory, Graph Theory, Combinatorics,
Computer Science Math.related topics,combinations of the above, etc. In general any kind of Concretely presented Mathematics which is Applicable fits to the scope of this journal.
Working Concretely and in Applicable Mathematics has become a main trend in many recent years,so we can understand better and deeper and solve the important problems of our real and scientific world.
"Journal of Concrete and Applicable Mathematics" is a peer- reviewed International Quarterly Journal.
We are calling for papers for possible publication. The contributor should send three copies of the contribution to the editor in-Chief typed in TEX, LATEX double spaced. [See: Instructions to Contributors]

Journal of Concrete and Applicable Mathematics(JCAAM)
ISSN:1548-5390 PRINT, 1559-176X ONLINE.
is published in January,April,July and October of each year by
EUDOXUS PRESS,LLC,
1424 Beaver Trail Drive,Cordova,TN38016,USA,
Tel.001-901-751-3553
anastassioug@yahoo.com
http://www.EudoxusPress.com.

Annual Subscription Current Prices:For USA and Canada,Institutional:Print $250,Electronic $220,Print and Electronic $310.Individual:Print $77,Electronic $60,Print &Electronic $110.For any other part of the world add $25 more to the above prices for Print.

Single article PDF file for individual $8.Single issue in PDF form for individual $25.
The journal carries page charges $8 per page of the pdf file of an article,payable upon acceptance of the article within one month and before publication.
No credit card payments.Only certified check,money order or international check in US dollars are acceptable.
Combination orders of any two from JoCAAA,JCAAM,JAFA receive 25% discount,all three receive 30% discount.

Copyright©2004 by Eudoxus Press,LLC all rights reserved.JCAAM is printed in USA.

JCAAM is reviewed and abstracted by AMS Mathematical Reviews,MATHSCI,and Zentralblaat MATH.

It is strictly prohibited the reproduction and transmission of any part of JCAAM and in any form and by any means without the written permission of the publisher.It is only allowed to educators to Xerox articles for educational purposes.The publisher assumes no responsibility for the content of published papers.
JCAAM IS A JOURNAL OF RAPID PUBLICATION

Editorial Board
Associate Editors

Editor in -Chief:
George Anastassiou
Department of Mathematical Sciences
The University Of Memphis
Memphis,TN 38152,USA
tel.901-678-3144,fax 901-678-2480
e-mail ganastss@memphis.edu
www.msci.memphis.edu/~ganastss/jcaam
Areas:Approximation Theory,
Probability,Moments,Wavelet,
Neural Networks,Inequalities,Fuzzyness.

Associate Editors:

1) Ravi Agarwal
Florida Institute of Technology
Applied Mathematics Program
150 W.University Blvd.
Melbourne,FL 32901,USA
agarwal@fit.edu
Differential Equations,Difference
Equations,inequalities

2) Shair Ahmad
University of Texas at San Antonio
Division of Math.& Stat.
San Antonio,TX 78249-0664,USA
SAHMAD@utsa.edu
Differential Equations,Mathematical
Biology

3) Drumi D.Bainov
Medical University of Sofia
P.O.Box 45,1504 Sofia,Bulgaria
drumibainov@yahoo.com
Differential Equations,Optimal Control,
Numerical Analysis,Approximation Theory

4) Carlo Bardaro
Dipartimento di Matematica & Informatica
Universita' di Perugia
Via Vanvitelli 1
06123 Perugia,ITALY
tel.+390755855034, +390755853822,
fax +390755855024
bardaro@unipg.it ,
bardaro@dipmat.unipg.it
Functional Analysis and Approximation Th.,

19) Rupert Lasser
Institut fur Biomathematik & Biomertie,GSF
-National Research Center for environment and
health
Ingolstaedter landstr.1
D-85764 Neuherberg,Germany
lasser@gsf.de
Orthogonal Polynomials,Fourier Analysis,
Mathematical Biology

20) Alexandru Lupas
University of Sibiu
Faculty of Sciences
Department of Mathematics
Str.I.Ratiu nr.7
2400-Sibiu,Romania
lupas@ulbsibiu.ro
Classical Analysis,Inequalities,
Special Functions,Umbral Calculus,
Approximation Th.,Numerical Analysis
and Methods

21) Ram N.Mohapatra
Department of Mathematics
University of Central Florida
Orlando,FL 32816-1364
tel.407-823-5080
ramm@mail.ucf.edu
Real and Complex analysis,Approximation Th.,
Fourier Analysis, Fuzzy Sets and Systems

22) Rainer Nagel
Arbeitsbereich Funktionalanalysis
Mathematisches Institut
Auf der Morgenstelle 10
D-72076 Tuebingen
Germany
tel.49-7071-2973242
fax 49-7071-294322
rana@fa.uni-tuebingen.de
evolution equations,semigroups,spectral th.,
positivity

23) Panos M.Pardalos
Center for Appl. Optimization
University of Florida
303 Weil Hall
P.O.Box 116595
Gainesville,FL 32611-6595

Summability,Signal Analysis,Integral Equations,
Measure Th.,Real Analysis

5) Francoise Bastin
Institute of Mathematics
University of Liege
4000 Liege
BELGIUM
f.bastin@ulg.ac.be
Functional Analysis,Wavelets

6) Paul L.Butzer
RWTH Aachen
Lehrstuhl A fur Mathematik
D-52056 Aachen
Germany
tel.0049/241/80-94627 office,
0049/241/72833 home,
fax 0049/241/80-92212
Butzer@rwth-aachen.de
Approximation Th.,Sampling Th.,Signals,
Semigroups of Operators,Fourier Analysis

7) Yeol Je Cho
Department of Mathematics Education
College of Education
Gyeongsang National University
Chinju 660-701
KOREA
tel.055-751-5673 Office,
055-755-3644 home,
fax 055-751-6117
yjcho@nongae.gsnu.ac.kr
Nonlinear operator Th.,Inequalities,
Geometry of Banach Spaces

8) Sever S.Dragomir
School of Communications and Informatics
Victoria University of Technology
PO Box 14428
Melbourne City M.C
Victoria 8001,Australia
tel 61 3 9688 4437,fax 61 3 9688 4050
sever.dragomir@vu.edu.au,
sever@sci.vu.edu.au
Math.Analysis,Inequalities,Approximation Th.,
Numerical Analysis, Geometry of Banach Spaces,
Information Th. and Coding

9) A.M.Fink
Department of Mathematics
Iowa State University
Ames,IA 50011-0001,USA
tel.352-392-9011
pardalos@ufl.edu
Optimization,Operations Research

24) Svetlozar T.Rachev
Dept.of Statistics and Applied Probability Program
University of California,Santa Barbara
CA 93106-3110,USA
tel.805-893-4869
rachev@pstat.ucsb.edu
AND
Chair of Econometrics and Statistics
School of Economics and Business Engineering
University of Karlsruhe
Kollegium am Schloss,Bau II,20.12,R210
Postfach 6980,D-76128,Karlsruhe,Germany
tel.011-49-721-608-7535
rachev@lsoe.uni-karlsruhe.de
Mathematical and Empirical Finance,
Applied Probability, Statistics and Econometrics

25) Paolo Emilio Ricci
Universita' degli Studi di Roma "La Sapienza"
Dipartimento di Matematica-Istituto "G.Castelnuovo"
P.le A.Moro,2-00185 Roma,ITALY
tel.++39 0649913201,fax ++39 0644701007
riccip@uniroma1.it,Paoloemilio.Ricci@uniroma1.it
Orthogonal Polynomials and Special functions,
Numerical Analysis, Transforms,Operational Calculus,
Differential and Difference equations

26) Cecil C.Rousseau
Department of Mathematical Sciences
The University of Memphis
Memphis,TN 38152,USA
tel.901-678-2490,fax 901-678-2480
ccrousse@memphis.edu
Combinatorics,Graph Th.,
Asymptotic Approximations,
Applications to Physics

27) Tomasz Rychlik
Institute of Mathematics
Polish Academy of Sciences
Chopina 12,87100 Torun, Poland
T.Rychlik@impan.gov.pl
Mathematical Statistics,Probabilistic Inequalities

28) Bl. Sendov
Institute of Mathematics and Informatics
Bulgarian Academy of Sciences
Sofia 1090,Bulgaria

tel.515-294-8150
fink@math.iastate.edu
Inequalities,Ordinary Differential Equations

10) Sorin Gal
Department of Mathematics
University of Oradea
Str.Armatei Romane 5
3700 Oradea,Romania
galso@uoradea.ro
Approximation Th.,Fuzzyness,Complex Analysis

11) Jerome A.Goldstein
Department of Mathematical Sciences
The University of Memphis,
Memphis,TN 38152,USA
tel.901-678-2484
jgoldste@memphis.edu
Partial Differential Equations,
Semigroups of Operators

12) Heiner H.Gonska
Department of Mathematics
University of Duisburg
Duisburg,D-47048
Germany
tel.0049-203-379-3542 office
gonska@informatik.uni-duisburg.de
Approximation Th.,Computer Aided Geometric Design

13) Dmitry Khavinson
Department of Mathematical Sciences
University of Arkansas
Fayetteville,AR 72701,USA
tel.(479)575-6331,fax(479)575-8630
dmitry@uark.edu
Potential Th.,Complex Analysis,Holomorphic PDE,Approximation Th.,Function Th.

14) Virginia S.Kiryakova
Institute of Mathematics and Informatics
Bulgarian Academy of Sciences
Sofia 1090,Bulgaria
virginia@diogenes.bg
Special Functions,Integral Transforms,
Fractional Calculus

15) Hans-Bernd Knoop
Institute of Mathematics
Gerhard Mercator University
D-47048 Duisburg
Germany
tel.0049-203-379-2676
bsendov@bas.bg
Approximation Th.,Geometry of Polynomials,
Image Compression

29) Igor Shevchuk
Faculty of Mathematics and Mechanics
National Taras Shevchenko
University of Kyiv
252017 Kyiv
UKRAINE
shevchuk@univ.kiev.ua
Approximation Theory

30) H.M.Srivastava
Department of Mathematics and Statistics
University of Victoria
Victoria,British Columbia V8W 3P4
Canada
tel.250-721-7455 office,250-477-6960 home,
fax 250-721-8962
harimsri@math.uvic.ca
Real and Complex Analysis,Fractional Calculus and Appl.,
Integral Equations and Transforms,Higher Transcendental
Functions and Appl.,q-Series and q-Polynomials,
Analytic Number Th.

31) Ferenc Szidarovszky
Dept.Systems and Industrial Engineering
The University of Arizona
Engineering Building,111
PO.Box 210020
Tucson,AZ 85721-0020,USA
szidar@sie.arizona.edu
Numerical Methods,Game Th.,Dynamic Systems,
Multicriteria Decision making,
Conflict Resolution,Applications
in Economics and Natural Resources Management

32) Gancho Tachev
Dept.of Mathematics
Univ.of Architecture,Civil Eng. and Geodesy
1 Hr.Smirnenski blvd
BG-1421 Sofia,Bulgaria
Approximation Theory

33) Manfred Tasche
Department of Mathematics
University of Rostock
D-18051 Rostock
Germany
manfred.tasche@mathematik.uni-rostock.de
Approximation Th.,Wavelet,Fourier Analysis,
Numerical Methods,Signal Processing,

knoop@math.uni-duisburg.de
Approximation Theory,Interpolation

16) Jerry Koliha
Dept. of Mathematics & Statistics
University of Melbourne
VIC 3010,Melbourne
Australia
koliha@unimelb.edu.au
Inequalities,Operator Theory,
Matrix Analysis,Generalized Inverses

17) Mustafa Kulenovic
Department of Mathematics
University of Rhode Island
Kingston,RI 02881,USA
kulenm@math.uri.edu
Differential and Difference Equations

18) Gerassimos Ladas
Department of Mathematics
University of Rhode Island
Kingston,RI 02881,USA
gladas@math.uri.edu
Differential and Difference Equations
Image Processing,Harmonic Analysis

34) Chris P.Tsokos
Department of Mathematics
University of South Florida
4202 E.Fowler Ave.,PHY 114
Tampa,FL 33620-5700,USA
profcpt@math.usf.edu,profcpt@chuma1.cas.usf.edu
Stochastic Systems,Biomathematics,
Environmental Systems,Reliability Th.

35) Lutz Volkmann
Lehrstuhl II fuer Mathematik
RWTH-Aachen
Templergraben 55
D-52062 Aachen
Germany
volkm@math2.rwth-aachen.de
Complex Analysis,Combinatorics,Graph Theory

JOURNAL OF CONCRETE AND APPLICABLE MATHEMATICS,VOL.3,NO.3,253-263,2005,COPYRIGHT 2005 EUDOXUS PRESS,LLC

On Meromorphic Approximation in the Space L_p

A.V. Krot,
Department of Mathematics and Mechanics
Belorussian State University
4 Scorina Ave., Minsk, 220080, Belarus

V.A. Prokhorov,
Department of Mathematics and Statistics
University of South Alabama
Mobile, Alabama 36688, U.S.A.
prokhoro@jaguar1.usouthal.edu
tel: (251) 460-6264, fax: (251) 460-7969

Abstract

Let $E_p(G)$, $1 \leq p < \infty$, be the Smirnov class of analytic functions on a bounded domain G containing 0. We assume that the boundary Γ of G consists of N closed analytic Jordan curves. For any nonnegative integers n and m denote by $\mathcal{M}^p_{n,m}$ the class of all meromorphic functions h on G that can be represented in the form $h = \alpha/\beta z^m$, where $\alpha \in E_p(G)$, β is a polynomial degree at most n, $\beta \not\equiv 0$. We investigate the problem of approximation of a function $f \in L_p(\Gamma)$ by functions in the class $\mathcal{M}^p_{n,m}$. A theorem giving necessary conditions for a function belonging to the class $\mathcal{M}^p_{n,m}$ to be an element of best approximation is proved. As a consequence we show that the denominators of the best meromorphic approximants are orthogonal polynomials with respect to a weight that varies with n.

AMS Classification: 41A20, 30E10, 30D50.
Key words: Meromorphic approximation, Best approximation, Orthogonal polynomial.

1 Necessary Conditions for an Element of Best Approximation

Let G be a bounded N-connected domain and let the boundary Γ of G consist of closed analytic Jordan curves. We assume that Γ is positively

oriented with respect to G and $0 \in G$. Let $1 \le p < \infty$. Given a pair (n, m) of nonnegative integers, we consider the following collection of meromorphic functions

$$\mathcal{M}^p_{n,m} = \mathcal{M}^p_{n,m}(G) = \{h = \alpha/\beta z^m,\ \alpha \in E_p(G), \beta \in \mathcal{P}_n, \beta \not\equiv 0\},$$

where $E_p(G)$ is the Smirnov class of analytic functions on G and $\mathcal{P}_n$ is the class of algebraic polynomials of degree at most n.

Denote by $L_p(\Gamma)$, $1 \le p < \infty$, the Lebesgue space of functions with integrable pth power. For $\varphi \in L_p(\Gamma)$ the norm of φ in $L_p(\Gamma)$ is

$$\|\varphi\|_p = \left(\int_\Gamma |\varphi(\xi)|^p |d\xi|\right)^{1/p}.$$

The Smirnov class $E_p(G)$, $1 \le p < \infty$, of analytic functions on G consists of the functions φ for which there is a sequence of domains G_k with rectifiable boundaries having the following properties

$$G_{k+1} \subset G_k, \quad \overline{G}_k \subset G, \quad \bigcup_k G_k = G$$

and

$$\sup_k \int_{\partial G_k} |\varphi(\xi)|^p |d\xi| < \infty.$$

Here and in what follows we consider the Smirnov class $E_p(G)$ as a subspace of the space $L_p(\Gamma)$. The condition

$$\int_\Gamma \frac{\varphi(\xi)\, d\xi}{\xi - z} = 0 \qquad \text{for all} \quad z \in \overline{\mathbf{C}} \setminus \overline{G} \tag{1}$$

is necessary and sufficient for a function $\varphi \in L_1(\Gamma)$ to be the boundary value of a function in the Smirnov class $E_1(G)$ (see [5] and [8] for more details about the classes $E_p(G)$).

Let f belong to the space $L_p(\Gamma)$, $1 \le p < \infty$. The *error in best approximation* of f in the space $L_p(\Gamma)$ by functions from the class $\mathcal{M}^p_{n,m}$ is denoted by $\Delta^p_{n,m}$:

$$\Delta^p_{n,m} = \Delta^p_{n,m}(f; G) = \inf_{h \in \mathcal{M}^p_{n,m}} \|f - h\|_p. \tag{2}$$

It is not difficult to prove that there exists a function $h_{n,m} \in \mathcal{M}^p_{n,m}$ for which the infimum in (2) is attained:

$$\Delta^p_{n,m} = ||f - h_{n,m}||_p$$

(see, for example, [6] for $p = \infty$).

We represent $h_{n,m}$ in the form $h_{n,m} = P/Qz^m$, where Q is a monic polynomial with all its zeros in G, $P \in E_p(G)$ and $P \neq 0$ at the zeros of Q.

From the relation

$$\Delta^p_{n,m} = \inf_{u \in E_p(G)} \|f - u/Q\xi^m\|_p, \tag{3}$$

we can assert with the help of duality relations (see [9], [10]; for the case when G is the unit disk see [3] and [4]) that there exists a unique function $\phi \in E_q(G)$, $||Q\xi^m\phi||_q = 1$, $1/p + 1/q = 1$, such that

$$Q(\xi)\xi^m\phi(\xi)(f - h_{n,m})(\xi)d\xi = \Delta^p_{n,m}|Q(\xi)\xi^m\phi(\xi)|^q\,|d\xi| \tag{4}$$

almost everywhere on Γ if $1 < p < \infty$, and

$$Q(\xi)\xi^m\phi(\xi)(f - h_{n,m})(\xi)d\xi = |(f - h_{n,m})(\xi)||d\xi| \tag{5}$$

almost everywhere on Γ if $p = 1$.

We now formulate a theorem giving necessary conditions for an element of $\mathcal{M}^p_{n,m}$ to be an element of best approximation to the function f in the space $L_p(\Gamma)$ in the class $\mathcal{M}^p_{n,m}$.

Theorem 1. *Suppose that $f \in L_p(\Gamma), 1 \leq p < \infty$, and $h_{n,m} = P/Qz^m$ is an element of best $L_p(\Gamma)$ approximation to f in the class $\mathcal{M}^p_{n,m}$. Then there exists a function $\varphi \in E_q(G)$ with $||Q^2\xi^m\varphi||_q = 1$, $1/p + 1/q = 1$, such that*

$$Q^2(\xi)\xi^m\varphi(\xi)(f - h_{n,m})(\xi)d\xi = \Delta^p_{n,m}|Q^2(\xi)\xi^m\varphi(\xi)|^q|d\xi|, \tag{6}$$

almost everywhere on Γ if $1 < p < \infty$, and

$$Q^2(\xi)\xi^m\varphi(\xi)(f - h_{n,m})(\xi)d\xi = |(f - h_{n,m})(\xi)||d\xi| \tag{7}$$

almost everywhere on Γ if $p = 1$. Moreover, for all $1 \leq p < \infty$ we can assert that $\deg Q = n$.

The occurrence of the constant 2 in the power of Q is a consequence of the fact that Q is constructed from free poles of the meromorphic approximants (compare with the factors in (6) and (7) that are connected with the "fixed part" of the element of best approximation).

In connection with Theorem 1 we single out paper [7], where an analog of (6) was obtained in the case when f is continuous on Γ and $p = \infty$.

This paper is organized as follows. The proof of Theorem 1 is presented in Section 2. In Section 3 questions concerning the orthogonal polynomials are investigated. Using equations (6) and (7) we prove that Q is the orthogonal polynomial with respect to a weight that varies with n.

2 Proof of Theorem 1

We now show that in the case when $1 < p < \infty$ the polynomial Q satisfies (6) on Γ. The case $p = 1$ can be treated in the same manner as $1 < p < \infty$. It will be assumed that $\Delta^p_{n,m} > 0$ since for $\Delta^p_{n,m} = 0$ the assertion is obvious.

Fix an arbitrary $u \in E_p(G)$. It is easy to see that we can represent u in the form $u = \alpha Q - \beta P$, where $\alpha \in E_p(G)$ and β is a polynomial with $\deg \beta \le n$. For any real ε, let

$$h_{n,m,\varepsilon} := \frac{P\omega + \varepsilon\alpha}{(Q\omega + \varepsilon\beta)\xi^m}, \tag{8}$$

where ω is an arbitrary monic polynomial of degree $d = n - \deg Q$ with all its zeros in G (for $d = 0$ we can take $\omega \equiv 1$). We remark that $h_{n,m,\varepsilon}$ belongs to the class $\mathcal{M}^p_{n,m}$. Assume that $|\varepsilon|$ is small enough so that $Q\omega + \varepsilon\beta \neq 0$ on Γ.

We choose a function $\phi \in E_q(G)$ with $||Q\xi^m\phi||_q = 1$, satisfying (4). Let us estimate

$$I_\varepsilon = \int_\Gamma Q(\xi)\xi^m\phi(\xi)(f - h_{n,m,\varepsilon})(\xi)d\xi.$$

Since

$$h_{n,m,\varepsilon} - h_{n,m} = \frac{\varepsilon u}{(Q\omega + \varepsilon\beta)Q\xi^m},$$

we have

$$\begin{aligned} I_\varepsilon &= \int_\Gamma Q(\xi)\xi^m\phi(\xi)\left(f(\xi) - h_{n,m}(\xi) - \frac{\varepsilon u(\xi)}{(Q\omega + \varepsilon\beta)(\xi)Q(\xi)\xi^m}\right)d\xi \\ &= \Delta^p_{n,m} - \varepsilon\int_\Gamma \frac{\phi(\xi)u(\xi)}{Q(\xi)\omega(\xi)}d\xi + o(\varepsilon) \quad \text{as} \quad \varepsilon \to 0. \end{aligned} \tag{9}$$

Next we claim that

$$||f - h_{n,m,\varepsilon}||_p = ||f - h_{n,m}||_p + o(\varepsilon) \quad \text{as} \quad \varepsilon \to 0. \tag{10}$$

Indeed, we have

$$||f - h_{n,m,\varepsilon}||_p = ||f - h_{n,m}||_p + \varepsilon\lambda + o(\varepsilon) \quad \text{as} \quad \varepsilon \to 0, \tag{11}$$

where

$$\lambda := -(\Delta^p_{n,m})^{1-p}\int_\Gamma |(f - h_{n,m})(\xi)|^{p-2}\,\mathrm{Re}\left((\overline{f - h_{n,m}})(\xi)\frac{u(\xi)}{(Q^2\omega)(\xi)}\right)|d\xi|.$$

Since, for any ε,

$$||f - h_{n,m,\varepsilon}||_p \ge ||f - h_{n,m}||_p,$$

from (11) we get that $\lambda = 0$. Thus (10) holds, and consequently

$$|I_\varepsilon| \le ||f - h_{n,m,\varepsilon}||_p ||Q\xi^m \phi||_q = ||f - h_{n,m}||_p + o(\varepsilon) = \Delta^p_{n,m} + o(\varepsilon) \quad (12)$$

as $\varepsilon \to 0$.

Letting $\varepsilon \to 0$ we obtain from (9) and (12) that

$$\int_\Gamma \frac{\phi(\xi)u(\xi)}{Q(\xi)\omega(\xi)} d\xi = 0. \quad (13)$$

Since ω is an arbitrary monic polynomial of degree d and u is an arbitrary function in $E_p(G)$, we can conclude from (13) that $d = 0, \deg Q = n$, and $\phi = Q\varphi$, where $\varphi \in E_q(G)$. In view of (4), this completes the proof of Theorem 1. •

3 A Connection with the Orthogonal Polynomials

3.1 G is the Unit Disk, $1 < p < \infty$

Our goal in this section is to prove that Q is a nonhermitian orthogonal polynomial for a weight that varies with n (cf. [1] and [7]).

Here and in the next subsections we assume that f is holomorphic on $\overline{\mathbf{C}} \setminus E$, where E is a compact set, $E \subset G$. Since f is holomorphic on Γ it follows from (6) and (7) that the function $Q^2\xi^m\varphi(f - h_{n,m})$ can be continued analytically across Γ (see, for example, [9], [10]). Let D be an arbitrary domain such that $E \subset D$ and $\overline{D} \subset G$. We assume that the boundary γ of D is positively oriented with respect to D and consists of finitely many disjoint contours.

We first consider the case when G is the unit disk with the center at 0. In this case the Smirnov class $E_p(G)$, $1 \le p < \infty$, coincides with the Hardy class $H_p(G)$ of analytic functions on G. The Hardy class $H_p(G)$ consists of the functions φ such that

$$\sup_{0<r<1} \int_0^{2\pi} |\varphi(re^{i\theta})|^p d\theta < \infty.$$

We say $\varphi \in H_p(G)$ is an outer function if for all $z \in G$

$$\ln|\varphi(z)| = \frac{1}{2\pi} \int_0^{2\pi} \ln|\varphi(e^{i\theta})| \, \frac{1 - |z|^2}{|e^{i\theta} - z|^2} d\theta.$$

A bounded nonvanishing analytic on G function φ is a singular function if $|\varphi(\xi)| = 1$ almost everywhere on Γ.

Each function φ in $H_p(G)$ has a factorization $\varphi = BS\mathrm{M}$ where B is a Blaschke product, S is a singular function, and M is an outer function. These factors are unique up to multiplicative constants of modulus one (see [3], [4]).

We remark that since the function $Q^2\xi^m\varphi(f - h_{n,m})$ can be continued analytically across Γ, φ has no a nonconstant singular factor, and φ has a finite number of zeros in G.

Let $n \geq 1$. Denote by B the Blaschke product

$$B(z) = \prod_{k=1}^{n} \frac{z - \xi_{k,n}}{1 - \overline{\xi}_{k,n} z}$$

constructed from the zeros of $Q(z) = \prod_{k=1}^{n}(z - \xi_{k,n})$. Set

$$w_n(z) = \prod_{k=1}^{n}(1 - \overline{\xi}_{k,n} z).$$

Denote by l_n the number of zeros of φ in G (counting multiplicity). For $l_n \geq 1$ let B_1 be the corresponding Blaschke product constructed from the zeros $\eta_{1,n}, \ldots, \eta_{l_n,n}$ of φ in G and let

$$\omega_n(z) = \prod_{k=1}^{l_n}(1 - \overline{\eta}_{k,n} z).$$

Set $\omega_n(z) \equiv 1$ and $B_1(z) \equiv 1$ if $l_n = 0$. We can now write

$$(Q^2\varphi)(z) = B^2(z) B_1(z) \psi^2(z),$$

where $\psi \in H_{2q}(G)$ and ψ is an outer function. Therefore, by (6) and (7),

$$B^2(\xi) B_1(\xi) \xi^m \psi^2(\xi)(f - h_{n,m})(\xi)\, d\xi = \Delta_{n,m}^p |\psi(\xi)|^{2q}\, |d\xi| \tag{14}$$

almost everywhere on Γ if $1 < p < \infty$, and

$$B^2(\xi) B_1(\xi) \xi^m \psi^2(\xi)(f - h_{n,m})(\xi) d\xi = |(f - h_{n,m})(\xi)||d\xi| \tag{15}$$

almost everywhere on Γ if $p = 1$.

We now show that for $1 < p < \infty$ the polynomial Q satisfies certain orthogonality relations with respect to a varying complex-valued measure and we find an integral expression for $f - h_{n,m}$.

Theorem 2. *Suppose that* $n \geq 1$ *and* $1 < p < \infty$. *Then the function* $B\xi^m\psi^{2-q}(f - h_{n,m})$ *can be extended analytically to* $\overline{\mathbf{C}} \setminus E$.

Moreover, the function $\psi^*(z) = \overline{\psi^q(\overline{z})}, z \in G$, *can be extended analytically to* $\overline{\mathbf{C}} \setminus E^{-1}$, *where* E^{-1} *is reflection of* E *in the unit circle, and satisfies the equation*

$$\frac{1}{2\pi}\int_\gamma \frac{(B^2B_1\xi^m\psi^{2-q}f)(\xi)d\xi}{z-\xi} = \Delta^p_{n,m}\psi^*\left(\frac{1}{z}\right)\frac{1}{z}, \quad z \in \overline{\mathbf{C}} \setminus \overline{D}. \tag{16}$$

Furthermore,

$$\int_\gamma \xi^\nu Q(\xi)\frac{\xi^m\psi^{2-q}(\xi)f(\xi)d\xi}{w_n^2(\xi)\omega_n(\xi)} = 0 \quad \textit{for} \quad \nu = 0, 1, \ldots, n + l_n - 1 \tag{17}$$

and, for $z \in \overline{\mathbf{C}} \setminus \overline{D}$,

$$(BT\xi^m\psi^{2-q}/w_n\omega_n)(z)(f - h_{n,m})(z) = \frac{1}{2\pi i}\int_\gamma \frac{(BT\xi^m\psi^{2-q}f)(\xi)\,d\xi}{w_n(\xi)\omega_n(\xi)(z-\xi)}, \tag{18}$$

where T *is any polynomial of degree at most* $n + l_n$.

Proof. Let $\Phi = B\xi^m\psi^2(f - h_{n,m})$. By (14),

$$\begin{aligned}\Phi(\xi)/\psi^q(\xi) &= B(\xi)\xi^m\psi^{2-q}(\xi)(f - h_{n,m})(\xi)\\ &= \Delta^p_{n,m}\frac{1}{B(\xi)B_1(\xi)}\overline{\psi^q(\xi)}\frac{1}{i\xi} = \Delta^p_{n,m}\frac{1}{B(\xi)B_1(\xi)}\psi^*\left(\frac{1}{\xi}\right)\frac{1}{i\xi}\end{aligned} \tag{19}$$

almost everywhere on Γ. Taking into account now that Φ can be continued analytically across Γ, ψ^q is an outer function, and the fact that the boundary values of Φ/ψ^q on Γ belong to $L_2(\Gamma)$, we can conclude that $\Phi/\psi^q \in H_2(U_r)$ for some $r > 0$, where $U_r = \{z : r < |z| < 1\}$. Thus, by (19), the functions $B\xi^m\psi^{2-q}(f - h_{n,m})$ and $B\xi^m\psi^{2-q}(f - h_{n,m})/w_n\omega_n$ can be extended analytically to $\overline{\mathbf{C}} \setminus E$ and the function $\psi^*(\xi)$ can be extended analytically to $\overline{\mathbf{C}} \setminus E^{-1}$. Moreover,

$$\frac{B\xi^m\psi^{2-q}(f - h_{n,m})}{w_n\omega_n}(z) = O\left[\frac{1}{z^{n+l_n+1}}\right] \quad \text{as} \quad z \to \infty. \tag{20}$$

Consequently,

$$\int_\gamma \xi^\nu Q(\xi)\frac{\xi^m\psi^{2-q}(\xi)f(\xi)d\xi}{w_n^2(\xi)\omega_n(\xi)} = 0 \quad \text{for} \quad \nu = 0, 1, \ldots, n + l_n - 1.$$

Let T be any polynomial of degree at most $n + l_n$. By the Cauchy formula,

$$(BT\xi^m\psi^{2-q}/w_n\omega_n)(z)(f-h_{n,m})(z) = \frac{1}{2\pi i}\int_\gamma \frac{(BT\xi^m\psi^{2-q}f)(\xi)\,d\xi}{w_n(\xi)\omega_n(\xi)(z-\xi)}, \quad z \in \overline{\mathbf{C}} \setminus \overline{D}.$$

In particular,

$$(B^2B_1\xi^m\psi^{2-q})(z)(f-h_{n,m})(z) = \frac{1}{2\pi i}\int_\gamma \frac{(B^2B_1\xi^m\psi^{2-q}f)(\xi)d\xi}{z-\xi}, \quad z \in \overline{\mathbf{C}} \setminus \overline{D}. \tag{21}$$

Then, by (19), we get

$$\frac{1}{2\pi}\int_\gamma \frac{(B^2B_1\xi^m\psi^{2-q}f)(\xi)d\xi}{z-\xi} = \Delta^p_{n,m}\psi^*\left(\frac{1}{z}\right)\frac{1}{z}, \quad z \in \overline{\mathbf{C}} \setminus \overline{D}. \qquad \bullet$$

3.2 The Case when G is the Unit Disk and $p = 1$

We now consider the case when $p = 1$. Let

$$v(z) = \exp\left(\frac{1}{4\pi}\int_\Gamma \frac{\xi+z}{\xi-z}\log\frac{|(f-h_n)(\xi)|}{\Delta^1_{n,m}}|d\xi|\right), \quad |z| < 1.$$

It is not hard to see that $v \in H_2(G)$, v is an outer function, $||v||_2 = 1$, and

$$|(f-h_{n,m})(\xi)| = \Delta^1_{n,m}|v(\xi)|^2$$

almost everywhere on Γ. So, by (15),

$$B^2(\xi)B_1(\xi)\xi^m\psi^2(\xi)(f-h_{n,m})(\xi)d\xi = \Delta^1_{n,m}|v(\xi)|^2|d\xi|$$

almost everywhere on Γ.

Applying the same arguments as in Subsection 3.1, we obtain the following theorem.

Theorem 3. *Suppose that $n \geq 1$ and $p = 1$. Then the function $B\xi^m\psi^2(f-h_{n,m})/v$ can be extended analytically to $\overline{\mathbf{C}} \setminus E$.*

Moreover, the function $v^*(z) = \overline{v(\overline{z})}, z \in G$, can be extended analytically to $\overline{\mathbf{C}} \setminus E^{-1}$, where E^{-1} is reflection of E in the unit circle, and satisfies the equation

$$\frac{1}{2\pi}\int_\gamma \frac{(B^2B_1\xi^m\psi^2f/v)(\xi)d\xi}{z-\xi} = \Delta^1_{n,m}v^*\left(\frac{1}{z}\right)\frac{1}{z}, \quad z \in \overline{\mathbf{C}} \setminus \overline{D}. \tag{22}$$

Furthermore,

$$\int_\gamma \xi^\nu Q(\xi)\frac{\xi^m\psi^2(\xi)f(\xi)d\xi}{w_n^2(\xi)\omega_n(\xi)v(\xi)} = 0 \quad \textit{for} \quad \nu = 0, 1, \ldots, n + l_n - 1 \tag{23}$$

and, for $z \in \overline{\mathbf{C}} \setminus \overline{D}$,

$$(BT\xi^m\psi^2/w_n\omega_n v)(z)(f - h_{n,m})(z) = \frac{1}{2\pi i}\int_\gamma \frac{(BT\xi^m\psi^2 f)(\xi)\, d\xi}{w_n(\xi)\omega_n(\xi)v(\xi)(z-\xi)}, \tag{24}$$

where T *is any polynomial of degree at most* $n + l_n$.

3.3 G is a Multiply Connected Domain

We consider the case when $1 < p < \infty$ and G is a bounded domain with the boundary Γ consisting of N disjoint closed analytic Jordan curves $\Gamma_k, k = 0, \ldots, N-1$. It will be assumed without loss of generality that $N \geq 2$. Assume that G lies interior to Γ_0. Let a_k be an arbitrary point lying inside of Γ_k, $k = 1, 2, \ldots, N-1$.

We represent $d\xi$ in the form $d\xi = S(\xi)|d\xi|$, $\xi \in \Gamma$. Let $\Phi_1 = Q^2\xi^m\varphi(f - h_{n,m})S$. By (6), $\Phi_1 \geq 0$ on Γ, and

$$\Phi_1(\xi)/(QS)(\xi) = \Delta_{n,m}^p\overline{Q(\xi)}|\xi^m|^q|Q(\xi)|^{2q-2}|\varphi(\xi)|^q/S(\xi)$$

almost everywhere on Γ. We assume that $\Phi_1/QS = Q\xi^m\varphi(f - h_{n,m}) \neq 0$ on Γ. Let $\chi = \mathrm{ind}_\Gamma(\Phi_1(\xi)/(QS)(\xi))$. It is easy to see that $\chi = N - 2 - n$. Let $n \geq N$, then $\chi \leq -2$.

Since $\chi \leq -2$, it follows from the results of the theory of boundary value problems for analytic functions (see, for example, [2]), that there exists a function g holomorphic on G and continuous on $\overline{G}$,

$$g(z) = \prod_{k=1}^{N-1}(z - a_k)\exp(-\Psi(z)), \quad z \in G,$$

where

$$\Psi(z) = \frac{1}{2\pi i}\int_\Gamma \frac{\log\left(\xi^{-\chi}\prod_{k=1}^{N-1}(\xi - a_k)\Phi_1(\xi)/(QS)(\xi)\right)}{\xi - z}\, d\xi,$$

such that the function $(gQ\xi^m\varphi)(z)(f - h_{n,m})(z)$ can be extended analytically to $\overline{\mathbf{C}} \setminus G$, and

$$(gQz^m\varphi)(z)(f - h_{n,m})(z) = O\left[\frac{1}{z^{|\chi|}}\right] \quad \text{as} \quad z \to \infty.$$

Consequently, Q satisfies the following orthogonality relations

$$\int_\Gamma \xi^\nu (gQ\xi^m \varphi f)(\xi) d\xi = 0, \qquad \nu = 0, \dots, |\chi| - 2. \tag{25}$$

Denote by T any polynomial of degree at most $|\chi| - 1$. By the Cauchy formula,

$$(TgQz^m\varphi)(z)(f - h_{n,m})(z) = \frac{1}{2\pi i} \int_\Gamma \frac{(TgQ\xi^m\varphi f)(\xi) d\xi}{z - \xi}, \quad z \in \overline{\mathbf{C}} \setminus \overline{G}.$$

Remark. The orthogonality equations (25) must be modified accordingly, if we drop the constraint that $Q\xi^m\varphi(f - h_{n,m}) \neq 0$ on Γ.

Acknowledgments. The authors would like to thank Professor E.B. Saff for his helpful comments and suggestions.

References

[1] L. Baratchart, V. A. Prokhorov, E. B. Saff, Best Meromorphic Approximation of Markov Functions on the Unit Circle, *Foundations of Computational Mathematics Journal* 1, 385-416 (2001).

[2] F. D. Gakhov, *Boundary Value Problems*, Pergamon Press, Oxford, 1966.

[3] J. B. Garnett, *Bounded Analytic Functions*, Academic Press, New York, 1981.

[4] P. Koosis, *Introduction to H^p-spaces*, Cambridge Univ. Press, Cambridge, 1980.

[5] I. I. Privalov, *Boundary Properties of Analytic Functions*, 2nd ed. GITTL, Moscow, 1950; German transl. VEB Deutscher Verlag Wiss., Berlin, 1956.

[6] V. A. Prokhorov, On a Theorem of Adamyan, Arov, and Kreĭn, *Mat. Sb.* 184, 89-104 (1993); English transl. in *Russian Acad. Sci. Sb. Math.* 78 (1994).

[7] V. A. Prokhorov and E. B. Saff, On Meromorphic Approximation, *Numer. Algorithms* 25, 305-321 (2000).

[8] G. Ts. Tumarkin and S. Ya. Khavinson, On the Definition of Analytic Functions of Class E_p in Multiply Connected Domains, *Uspekhi Mat. Nauk* 13, 201–206 (1958) (Russian).

[9] G. Ts. Tumarkin and S. Ya. Khavinson, Extremal Problems for Certain Classes of Analytic Functions on Finitely Connected Domains, *Mat. Sb.* 36(78), 445–478 (1955) (Russian).

[10] G. Ts. Tumarkin and S. Ya. Khavinson, Investigation of the Properties of Extremal Functions with the Help of Duality Relations in Extremal Problems for Classes of Analytic Functions on Multiply Connected Domains, *Mat. Sb.* 46(88), 195–228 (1958) (Russian).

JOURNAL OF CONCRETE AND APPLICABLE MATHEMATICS,VOL.3,NO.3,265-281,2005,COPYRIGHT 2005 EUDOXUS PRESS,LLC

Persistence and Global Stability for Pure-delay Type Nonautonomous Lotka-Volterra Differential Systems

Yoshiaki Muroya

Department of Mathematical Sciences, Waseda University
3-4-1 Ohkubo Shinjuku-ku, Tokyo, 169-8555 JAPAN.
Tel: +81-3-5286-3351
Fax: +81-3-3232-6795
Email: ymuroya@waseda.jp

Abstract

Consider the persistence and the global asymptotic stability of models governed by the following pure-delay type nonautonomous Lotka-Volterra differential system:

$$\begin{cases} \dfrac{dx_i(t)}{dt} = x_i(t)\{c_i(t) - \displaystyle\sum_{j=1}^{n}\sum_{l=0}^{m} a_{ij}^l(t)x_j(t-\tau_l)\},\ t \geq t_0,\ 1 \leq i \leq n, \\ x_i(t) = \phi_i(t) \geq 0,\ t \leq t_0,\ \textit{and}\ \phi_i(t_0) > 0,\ 1 \leq i \leq n, \end{cases}$$

where each $\phi_i(t)$ is a continuous function for $t \leq t_0$, each $c_i(t)$ and $a_{ij}^l(t)$ are bounded continuous functions on $[t_0, +\infty)$ and

$$\inf_{t \geq t_0}(\sum_{l=0}^{m} a_{ii}^l(t)) > 0,\ a_{ij}^l(t) \geq 0,\ i \leq j \leq n,\ 1 \leq i \leq n,\ \textit{and}\ \tau_l \geq 0,\ 0 \leq l \leq m.$$

In this paper, for the above system of pure-delay type, by improving the former work (*Communications on Applied Nonlinear Analysis* **9**, 31-45) which extended the averaged condition offered by S. Ahmad and A. C. Lazer (2000, *Nonlinear Analysis* **40**, 37-49), we offer conditions of persistence which are independent of time delays, and considering a continuous Lyapunov-like (nonnegative and non-differentiable) function to the above pure-delay type system, we establish sufficient conditions of global asymptotic stability.

AMS Subject Classifications : 34K20, 34K60
Key words : persistence ; global asymptotic stability ; pure-delay type nonautonomous Lotka-Volterra differential system.

1 Introduction

Consider conditions of the persistence and global asymptotic stability of the following pure-delay type nonautonomous Lotka-Volterra delay differential system.

$$\begin{cases} \dfrac{dx_i(t)}{dt} = x_i(t)\{c_i(t) - \displaystyle\sum_{j=1}^{n}\sum_{l=0}^{m} a_{ij}^l(t)x_j(t-\tau_l)\},\ t \geq t_0, \\ x_i(t) = \phi_i(t) \geq 0,\ t \leq t_0,\ \textit{and}\ \phi_i(t_0) > 0,\ 1 \leq i \leq n, \end{cases} \tag{1}$$

where each $\phi_i(t)$ is a continuous function for $t \le t_0$, each $c_i(t)$ and $a_{ij}^l(t)$ are bounded continuous functions on $[t_0, +\infty)$ and

$$\inf_{t\ge t_0}(\sum_{l=0}^{m} a_{ii}^l(t)) > 0,\ a_{ij}^l(t) \ge 0,\ i \le j \le n,\ 1 \le i \le n,\ \tau_l \ge 0,\ 0 \le l \le m. \tag{2}$$

Recently, in [11], Xu and Chen have established conditions for the permanence of system and using a Lyapunov functional, they obtained conditions for the global asymptotic stability of the time-dependent pure-delay-type Lotka-Volterra predator-prey model of three species.

In this paper, for the above system of pure-delay type, by applying the former work of Muroya [7] which extended the averaged condition offered by Ahmad and Lazer [2], we offer conditions of persistence which are independent of time delays, and considering a continuous Lyapunov-like (nonnegative and non-differentiable) function to the above pure-delay type system, we establish sufficient conditions of global asymptotic stability which extend the results of Zhen and Ma [12] and are a different type from those of Xu and Chen [11].

Let

$$\begin{cases} \bar{a}_i(t) = \sum_{l=0}^{m} a_{ii}^l(t),\ \bar{a}_{ii}^l(t) \equiv 0,\ \bar{a}_{ij}^l(t) = a_{ij}^l(t),\ j \ne i,\ 0 \le l \le m, \\ \bar{a}_{iL} = \inf_{t\ge t_0}(\sum_{l=0}^{m} a_{ii}^l(t)),\ \bar{a}_{iM} = \sup_{t\ge t_0}(\sum_{l=0}^{m} a_{ii}^l(t)),\ c_{iL} = \inf_{t\ge t_0} c_i(t),\ c_{iM} = \sup_{t\ge t_0} c_i(t), \\ \bar{a}_{ij}^{l-}(t) = \min(0, \bar{a}_{ij}^l(t)),\ \bar{a}_{ij}^{l+}(t) = \max(0, \bar{a}_{ij}^l(t)), \\ \bar{b}_{ijL}^{-} = \inf_{t\ge t_0}(\sum_{l=0}^{m} \bar{a}_{ij}^{l-}(t)),\ and\ \bar{b}_{ijM}^{+} = \sup_{t\ge t_0}(\sum_{l=0}^{m} \bar{a}_{ij}^{l+}(t)),\ 1 \le i, j \le n, \\ m[c_i] = \lim_{t\to\infty} \inf\{\ \frac{1}{t_2 - t_1}\int_{t_1}^{t_2} c_i(s)ds \mid t_0 \le t_1 < t_2\ and\ t_2 - t_1 \ge t\}, \\ and \\ M[c_i] = \lim_{t\to\infty} \sup\{\ \frac{1}{t_2 - t_1}\int_{t_1}^{t_2} c_i(s)ds \mid t_0 \le t_1 < t_2\ and\ t_2 - t_1 \ge t\},\ 1 \le i \le n. \end{cases} \tag{3}$$

Note that $\bar{b}_{iiL}^{-} = \bar{b}_{iiM}^{+} = 0$ and $c_{iL} \le m[c_i] \le M[c_i] \le c_{iM},\ 1 \le i \le n$.

First, applying the results of Muroya [7] to this system, we improve sufficient conditions for the persistence of system.

Then, the newly extended averaged condition is as follows:
For any $x_i \ge 0,\ 1 \le i \le n$ such that

$$M[c_i] \ge \bar{a}_{iL}x_i + \sum_{j=1}^{n} \bar{b}_{ijL}^{-}x_j,\ 1 \le i \le n, \tag{4}$$

it holds that

$$m[c_i] > \sum_{j\ne i} \bar{b}_{ijM}^{+}x_j,\ 1 \le i \le n. \tag{5}$$

Let

$$\begin{cases} \bar{A}_L = \mathrm{diag}(\bar{a}_{1L}, \bar{a}_{2L}, \cdots, \bar{a}_{nL}),\ \bar{B}_L^- = [\bar{b}_{ijL}^-],\ \bar{B}_M^+ = [\bar{b}_{ijM}^+],\ \textit{are } n \times n \textit{ matrices}, \\ \textit{and } \underline{\boldsymbol{c}} = [m[c_i]] \textit{ and } \bar{\boldsymbol{c}} = [M[c_i]] \textit{ are } n\textit{−dimensional vectors}. \end{cases} \tag{6}$$

Then, the condition (4)-(5) is equivalent to the following:
For any n-dimensional vector $\boldsymbol{x} = [x_i] \geq \mathbf{0}$ such that

$$\bar{\boldsymbol{c}} \geq (\bar{A}_L + \bar{B}_L^-)\boldsymbol{x}, \tag{7}$$

it holds that

$$\underline{\boldsymbol{c}} > \bar{B}_M^+ \boldsymbol{x}. \tag{8}$$

Assume that

$$\begin{cases} \bar{A}_L + \bar{B}_L^- \textit{ is an } M\textit{−matrix},\ (\bar{A}_L + \bar{B}_L^-)^{-1}\bar{\boldsymbol{c}} > \mathbf{0} \\ \textit{and } \underline{\boldsymbol{c}} > \bar{B}_M^+(\bar{A}_L + \bar{B}_L^-)^{-1}\bar{\boldsymbol{c}}. \end{cases} \tag{9}$$

Then, $(\bar{A}_L + \bar{B}_L^-)^{-1} \geq \mathbf{0}$ and Eq.(7) implies that

$$(\bar{A}_L + \bar{B}_L^-)^{-1}\bar{\boldsymbol{c}} \geq \boldsymbol{x},$$

and from $\underline{\boldsymbol{c}} > \bar{B}_M^+(\bar{A}_L + \bar{B}_L^-)^{-1}\bar{\boldsymbol{c}}$ and $\bar{B}_M^+ \geq \mathbf{0}$, we have that

$$\underline{\boldsymbol{c}} > \bar{B}_M^+(\bar{A}_L + \bar{B}_L^-)^{-1}\bar{\boldsymbol{c}} \geq \bar{B}_M^+ \boldsymbol{x},$$

which implies Eq.(8).
Thus, the newly extended averaged condition (4)-(5) is satisfied.

Put

$$\begin{cases} \bar{\tau} = \max\limits_{0 \leq l \leq m} \tau_l,\ \bar{\tau}_i = \max\{\tau_l \mid a_{ii}^l \neq 0,\ 0 \leq l \leq m\},\ 1 \leq i \leq n, \\ \tilde{x}_1 = \frac{c_{1M}}{\bar{a}_{1L}},\ \bar{x}_1 = \tilde{x}_1 \exp(c_{1M}\bar{\tau}_1), \\ \tilde{x}_i = (c_{iM} - \sum\limits_{j=1}^{i-1} \bar{b}_{ijL}^- \bar{x}_j)/\bar{a}_{iL},\ \bar{x}_i = \tilde{x}_i \exp\{(c_{iM} - \sum\limits_{j=1}^{i-1} \bar{b}_{ijL}^- \bar{x}_j)\bar{\tau}_i\},\ 2 \leq i \leq n-1, \\ \tilde{x}_n = (c_{nM} - \sum\limits_{j=1}^{n-1} \bar{b}_{njL}^- \bar{x}_j)/\bar{a}_{nL},\ \bar{x}_n = \tilde{x}_n \exp\{(c_{nM} - \sum\limits_{j=1}^{n-1} \bar{b}_{njL}^- \bar{x}_j)\bar{\tau}_n\}, \end{cases} \tag{10}$$

and assume

$$c_{iM} - \sum_{j=1}^{i-1} \bar{b}_{ijL}^- \bar{x}_j > 0,\ 1 \leq i \leq n. \tag{11}$$

Then,

$$\bar{x}_i \geq \tilde{x}_i > 0,\ 1 \leq i \leq n.$$

Note that if $\bar{A}_L + B_L^-$ is an M-matrix and $(\bar{A}_L + \bar{B}_L^-)^{-1}\bar{\boldsymbol{c}} > \mathbf{0}$, then for n-dimensional vector $\bar{\boldsymbol{x}} = [\bar{x}_i]$ and $\boldsymbol{c}_M = [c_{iM}]$, we have that

$$\bar{\boldsymbol{x}} \geq (\bar{A}_L + \bar{B}_L^-)^{-1}\boldsymbol{c}_M \geq (\bar{A}_L + \bar{B}_L^-)^{-1}\bar{\boldsymbol{c}} > \mathbf{0}. \tag{12}$$

Thus, Eq.(9) implies Eqs.(11) and (12).

For any two solutions $\{x_i(t)\}_{i=1}^n$ and $\{y_i(t)\}_{i=1}^n$ of the system (1)-(2), we have the following equation which will be useful to obtain sufficient conditions for the global asymptotic stability of system (see Lemma 8):

$$\left\{\begin{array}{l} \frac{d}{dt}\{\ln(x_i(t)/y_i(t))\} = -\bar{a}_i(t)(x_i(t)-y_i(t)) - \sum_{j=1}^{n}\sum_{l=0}^{m}\bar{a}_{ij}^l(t)(x_j(t-\tau_l)-y_j(t-\tau_l)) \\ \qquad + \sum_{l=0}^{m} a_{ii}^l(t)\int_{t-\tau_l}^{t}[\{c_i(s) - \sum_{j=1}^{n}\sum_{k=0}^{m} a_{ij}^k(s)y_j(s-\tau_k)\}(x_i(s)-y_i(s)) \\ \qquad - \sum_{j=1}^{n}\sum_{k=0}^{m} x_i(s)a_{ij}^k(s)(x_j(s-\tau_k)-y_j(s-\tau_k))]ds, \\ \qquad\qquad\qquad\qquad t \geq t_0+\bar{\tau},\ 1\leq i\leq n. \end{array}\right. \tag{13}$$

Put an $n\times n$ matrix as

$$\bar{\tilde{A}} = [\bar{\tilde{a}}_{ij}],$$

where

$$\left\{\begin{array}{l} a_{iiM}^l = \sup_{t\geq t_0} a_{ii}^l(t), \\ \bar{\tilde{a}}_{ii} = \bar{a}_{iL} - (\sum_{l=0}^{m} a_{iiM}^l\tau_l)\{\bar{a}_{iM}\bar{x}_i + \max(-c_{iL} + (\bar{a}_{iM}\bar{x}_i + \sum_{j=1}^{n}\bar{b}_{ijM}^+\bar{x}_j), c_{iM} - \sum_{j=1}^{n}\bar{b}_{ijL}^-\bar{x}_j)\}, \\ \bar{\tilde{a}}_{ij} = (-b_{ijL}^- + b_{ijM}^+)\{1 + (\sum_{l=0}^{m} a_{iiM}^l\tau_l)\bar{x}_i\},\ j\neq i,\ 1\leq i\leq n. \end{array}\right. \tag{14}$$

We shall establish the following extension of the Ahmad and Lazer's results in [2] and Muroya [7] to the system (1)-(2).

Theorem 1. (*Cf. Ahmad and Lazer [2] and Muroya [7]*)
For the system (1)-(2), *assume Eq.*(9) *and suppose that there exists a nonempty subset* $Q\in\{1,2,\cdots,n\}$ *such that*

$$c_{iL} - \sum_{j\notin Q} b_{ijM}^+\bar{x}_j > 0,\ for\ any\ i\in Q. \tag{15}$$

Then, the system is persistent for solutions, that is,

$$0 < \inf_{t\geq t_0} x_i(t) \leq \sup_{t\geq t_0} x_i(t) < +\infty,\ 1\leq i\leq n. \tag{16}$$

Moreover, if

$$\bar{\tilde{A}}\ is\ an\ M-matrix, \tag{17}$$

then for any two solutions $x_i(t)$ *and* $y_i(t)$, $1\leq i\leq n$, *it holds*

$$\lim_{t\to\infty}(x_i(t)-y_i(t)) = 0,\ 1\leq i\leq n. \tag{18}$$

In particular, if there exists a positive equilibrium $x^* = (x_1^*, x_2^*, \cdots, x_n^*)$ *of the system* (1)-(2)*, then in Eq.*(14)*, we can take* $\bar{\bar{a}}_{ii},\ 1 \le i \le n$ *as*

$$\bar{\bar{a}}_{ii} = \bar{a}_{iL} - (\sum_{l=0}^{n} a_{iiM}^l \tau_l)\bar{a}_{iM}\bar{x}_i,\ 1 \le i \le n.$$

Note that if

$$\bar{A}_L - (-B_L^- + B_M^+)\ is\ an\ M - matrix,$$

then Eq.(17) holds for sufficiently small delays such that $\sum_{l=0}^{m} a_{iiM}^l \tau_l,\ 1 \le i \le n$ are sufficiently small.

The organization of this paper is as follows. In Section 2, using the same techniques in Ahmad and Lazer [2] and Muroya [7], we prove that Eqs.(9) and (15) $\Rightarrow$ Eq.(16), and Eqs.(16) and (17) $\Rightarrow$ Eq.(18).

2 Conditions of persistence and global asymptotic stability

Consider the persistence and the global asymptotic stability of models governed by a pure-delay-type nonautonomous Lotka-Volterra delay differential system (1)-(2).

We have a lemma.

Lemma 1. *For the system* (1)-(2) *and* $1 \le i \le n$*,*

$$x_i(t) = x_i(t_0)\exp(\int_{t_0}^{t}\{c_i(s) - \sum_{j=1}^{n}\sum_{l=0}^{m} a_{ij}^l(s)x_j(s-\tau_l)\}ds),\ t \ge t_0, \tag{19}$$

and every solutions $x_i(t),\ 1 \le i \le n$*, exist and remain positive for all* $t \ge t_0$*.*

Proof. From Eq.(1), we have for any $t \ge t_0$,

$$\frac{d}{dt}\{\frac{1}{x_i(t)}\exp(\int_{t_0}^{t}\{c_i(s) - \sum_{j=1}^{n}\sum_{l=0}^{m} a_{ij}^l(s)x_j(s-\tau_l)\}ds)\} = 0.$$

Thus, integrating both sides with respect to t on $[t_0, t]$, one obtains Eq.(19), from which we get the conclusion. □

We have the following lemma which is an extension of Theorem 2.1 in Xu and Chen [11] for a time-dependent pure-delay-type Lotka-Volterra-prey model of three species.

Lemma 2. *For Eqs.*(3) *and* (10)*, assume Eq.*(11)*.*
Then, any solutions $x_i(t),\ 1 \le i \le n$ *of system* (1)-(2)*, are bounded above and*

$$\limsup_{t\to\infty} x_i(t) \le \bar{x}_i,\ 1 \le i \le n.$$

Proof. Let D^- denote the lower left Dini-derivative. If for some $t \ge t_0$, $D^- x_1(t) \ge 0$, then there exists a nonnegative integer $\bar{l}_{1t}$ such that $0 \le \bar{l}_{1t} \le m$ and $x_1(t-\tau_{\bar{l}_{1t}}) \le \frac{c_{1M}}{\bar{a}_{1L}} = \tilde{x}_1$.
Because, if

$$\min_{0\le l\le m} x_1(t-\tau_l) > \frac{c_{1M}}{\bar{a}_{1L}},$$

then

$$c_1(t) - \sum_{j=1}^{n}\sum_{l=0}^{m} a_{1j}^l(t)x_j(t-\tau_l) \le c_{1M} - \bar{a}_{1L}(\min_{0\le l\le m} x_1(t-\tau_l)) < 0,$$

which implies $D^- x_1(t) < 0$, by Eq.(1).
Therefore, by Eq.(19), $x_1(t) \le x_1(t-\tau_{\bar{l}_{1t}})\exp(c_{1M}\tau_{\bar{l}_{1t}}) \le \bar{x}_1$.
Thus, if $x_1(t) > \bar{x}_1$ for some $t \ge t_0$, then we have

$$D^- x_1(t) < 0.$$

Now, let us consider the case that $x_1(t)$ is eventually decreasing and bounded below by $\bar{x}_1$. Then, $\lim_{t\to\infty} x_1(t)$ exists. Set $\beta = \lim_{t\to\infty} x_1(t) - \bar{x}_1 \ge 0$.
We will show that $\beta = 0$.
Indeed, suppose $\beta > 0$. Let take any positive constant η. Then, there exists $\tilde{t}_0 \ge t_0$ such that

$$\beta \le x_1(s) - \bar{x}_1 \le \beta + \eta,\ for\ s \ge \tilde{t}_0,$$

since $x_1(t) - \bar{x}_1$ eventually decreases to β. Thus, we have

$$\begin{aligned} D^- x_1(t) &\le x_1(t)\{c_{1M} - \sum_{j=1}^{n}\sum_{l=0}^{m} a_{11}^l(t)x_1(t-\tau_l)\} \\ &\le x_1(t)(-\bar{a}_{1L}\beta),\ for\ t \ge \tilde{t}_1 \equiv \tilde{t}_0 + \bar{\tau}. \end{aligned}$$

Therefore, we have

$$x_1(t) \le x_1(\tilde{t}_1)\exp\{-\beta\int_{\tilde{t}_1}^{t} \bar{a}_{1L}ds\},$$

which in turn implies, due to $\int_{\tilde{t}_1}^{\infty} \bar{a}_{1L}ds = +\infty$, $\lim_{t\to\infty} x_1(t) = 0$.
This contradicts $x_1(t) \ge \bar{x}_1 + \beta > 0$. Thus, $\lim_{t\to\infty} x_1(t) = \bar{x}_1$.
Hence, we have

$$\limsup_{t\to\infty} x_1(t) \le \bar{x}_1.$$

Then, for any fixed positive constant ϵ, there exists a constant $\bar{t}_1 \ge \bar{t}_0 = t_0$ such that $x_1(t) \le \bar{x}_1$, for any $t \ge \bar{t}_1 - \bar{\tau}$.

Next, for some $2 \le i \le n$, suppose inductively that for any fixed positive constant ϵ, there exists a constant $\bar{t}_{i-1} \ge \bar{t}_{i-2}$ such that

$$x_j(t) \le \bar{x}_j + \epsilon, \ \textit{for any } t \ge \bar{t}_{i-1} - \bar{\tau}, \ 1 \le j \le i-1.$$

If for some $t \ge \bar{t}_i$, $D^- x_i(t) \ge 0$, then there exists a nonnegative integer $\underline{l}_{it}$ such that $0 \le \underline{l}_{it} \le m$ and

$$x_i(t - \tau_{\underline{l}_{it}}) \le \{c_{iM} - \sum_{j=1}^{i-1} \bar{b}^-_{ijL}(\bar{x}_j + \epsilon)\}/\bar{a}_{iL} \le \tilde{x}_i + \{(-\sum_{j=1}^{i-1} \bar{b}^-_{ijL})/\bar{a}_{iL}\}\epsilon.$$

Because, if

$$\min_{0 \le l \le m} x_i(t - \tau_l) > (c_{iM} - \sum_{j=1}^{n} \bar{b}^-_{ijL}(\bar{x}_j + \epsilon))/\bar{a}_{iL},$$

then

$$c_i(t) - \sum_{j=1}^{n}\sum_{l=0}^{m} a^l_{ij}(t) x_j(t - \tau_l) \le c_{iM} - \sum_{j=1}^{i-1} \bar{b}^-_{ijL}(\bar{x}_j + \epsilon) - \bar{a}_{iL}(\min_{0 \le l \le m}(x_i(t - \tau_l)) < 0,$$

which implies $D^- x_i(t) < 0$, by Eq.(1).
Therefore, by Eq.(19),

$$\begin{aligned} x_i(t) &\le x_i(t - \tau_{\underline{l}_{it}}) \exp\{(c_{iM} - \sum_{j=1}^{i-1} \bar{b}^-_{ijL}(\bar{x}_j + \epsilon))\tau_{\underline{l}_{it}}\} \\ &\le \bar{x}^\epsilon_i \equiv [\bar{x}_i + \{(-\sum_{j=1}^{i-1} \bar{b}^-_{ijL})/\bar{a}_{iM}\}\epsilon \exp\{(c_{iM} - \sum_{j=1}^{i-1} \bar{b}^-_{ijL}\bar{x}_j)\tau_{\underline{l}_{it}}\}] \\ &\quad \times \exp\{(-\sum_{j=1}^{i-1} \bar{b}^-_{ijL})\epsilon\tau_{\underline{l}_{it}}\}. \end{aligned}$$

Thus, if there exists a constant $\tilde{t}_i \ge \bar{t}_{i-1}$ such that $x_i(t) > \bar{x}^\epsilon_i$ for some $t \ge \tilde{t}_i$, then

$$D^- x_i(t) < 0.$$

If $x_i(t)$ is eventually decreasing and bounded below by $\bar{x}^\epsilon_i$. Then, as similar to the above discussions of $i = 1$, we see $\lim_{t \to \infty} x_i(t) = \bar{x}^\epsilon_i$.
Since $\epsilon > 0$ is any positive constant, we have that by inductions of $i = 1, 2, \cdots, n$,

$$\limsup_{t \to \infty} x_i(t) \le \bar{x}_i, \ 1 \le i \le n.$$

This completes the proof. □

The following lemma is an improved result to Lemma 2.2 in Ahmad and Lazer [2].

Lemma 3. *For Eqs.*(3) *and* (10)*, assume Eqs.*(11) *and* (15)*. Then, for solutions* $x_i(t), \ 1 \le i \le n$ *of system* (1)-(2)*,*

$$\liminf_{t \to \infty} \sum_{i \in Q} x_i(t) > 0. \tag{20}$$

Proof. By assumptions to the system (1)-(2), there exist positive constants $\underline{\gamma},\ \bar{b}_l,\ 0 \le l \le m$ such that for $i \in Q$,

$$c_{iL} - \sum_{j \notin Q} \bar{b}^{+}_{ijM} \bar{x}_j \ge \underline{\gamma},\ \textit{and}\ \bar{a}_{iM},\ \bar{a}^{l}_{ijM} \le \bar{b}_l,\ j \in Q,\ 0 \le l \le m.$$

By Eq.(1), it follows that for $i \in Q$,

$$\begin{aligned} x_i'(t) &\ge x_i(t)\{c_i(t) - \sum_{j \notin Q} \sum_{l=0}^{m} \bar{a}^{l+}_{ijM} x_j(t-\tau_l) - \bar{a}_i(t)x_i(t) - \sum_{j \in Q} \sum_{l=0}^{m} \bar{a}^{l}_{ij}(t) x_j(t-\tau_l)\} \\ &\ge x_i(t)\{\underline{\gamma} - \sum_{l=0}^{m} \bar{b}_l \sum_{j \in Q} x_j(t-\tau_l)\}. \end{aligned}$$

This shows that if

$$V(t) = \sum_{j \in Q} x_j(t),$$

then

$$V'(t) \ge V(t)\{\underline{\gamma} - \sum_{l=0}^{m} \bar{b}_l V(t-\tau_l)\}. \tag{21}$$

Now, suppose that $\liminf_{t \to \infty} V(t) = 0$. Then, there exists a sequence $\{t_p\}_{p=1}^{\infty}$ such that

$$V'(t_p) \le 0,\ \textit{and} \lim_{p \to \infty} V(t_p) = 0.$$

Since $V(t) > 0$ and for $V^* = \underline{\gamma}/(\sum_{l=0}^{m} \bar{b}_l) > 0$,

$$V'(t) \ge V(t) \sum_{l=0}^{m} \bar{b}_l (V^* - V(t-\tau_l)),$$

it holds that for each $p \ge 1$, there exists a $l_p \in \{0, 1, 2, \cdots, m\}$ such that

$$V(t_p - \tau_{l_p}) \ge V^*.$$

Similar to Eq.(19), it follows from Eq.(21) that

$$V(t_p) \ge V(t_p - \tau_{l_p}) \exp(\int_{t_p - \tau_{l_p}}^{t_p} \{\underline{\gamma} - \sum_{l=0}^{m} \bar{b}_l V(s-\tau_l)\} ds).$$

By Lemma 2 and assumptions, there is a positive constant $\bar{V}$ such that for $V(t) \le \bar{V},\ t \ge t_0$ and for $\bar{\tau} = \max_{0 \le l \le m} \tau_l$, we have that

$$V(t_p) \ge \beta \equiv V^* \exp(\{\underline{\gamma} - \sum_{l=0}^{m} \bar{b}_l \bar{V}\}\bar{\tau}) > 0,\ p \ge 1,$$

which is a contradiction.
Therefore,

$$\liminf_{t\to\infty} V(t) > 0,$$

and hence, Eq.(20) holds. □

From Lemma 3, we easily obtain the following Lemmas 4-6 (see Lemmas 2.3-2.5 in Ahmad and Lazer [2]).

Lemma 4. *For Eqs.*(3) *and* (10)*, assume Eqs.*(11) *and* (15) *and suppose that Eq.*(16) *does not hold. Then for solutions* $x_i(t),\ 1 \le i \le n$ *of system* (1)-(2)*, there exists a maximal nonempty subset* J *of* $\{1, 2, \cdots, n\}$ *such that*

$$J \neq \{1, 2, \cdots, n\}$$

and

$$\inf_{t\ge t_0} \max\{\ x_j(t) \mid j \in J\} = 0.$$

Lemma 5. *For Eqs.*(3) *and* (10)*, assume Eqs.*(11) *and* (15) *and suppose that Eq.*(16) *does not hold. Let* $x_i(t),\ 1 \le i \le n$ *and* J *be as in* Lemma 4 *and put*

$$\min\{\ x_j(t_0) \mid j \in J\} = \delta > 0.$$

Then, there exist sequences $\{s_p\}_{p=1}^{\infty}$ *and* $\{t_p\}_{p=1}^{\infty}$ *such that for any* $p \ge 1$*,*

$$\begin{cases} t_0 \le s_p < t_p, \\ t_p - s_p \ge p, \\ \max\{\ x_j(t) \mid j \in J,\ s_p \le t \le t_p\} = \delta/p, \end{cases} \tag{22}$$

and there exists $j_p \in J$ *such that*

$$x_{j_p}(s_p) = \delta/p. \tag{23}$$

The following lemma is a little improved version of Lemma 2.5 in Ahmad and Lazer [2] (see Lemma 2.6 in Muroya [7]).

Lemma 6. *For Eqs.*(3) *and* (10)*, assume Eqs.*(11) *and* (15) *and suppose that Eq.*(16) *does not hold. Let* $x_i(t),\ 1 \le i \le n,\ J$*, and the sequences* $\{s_p\}_{p=1}^{\infty},\ \{t_p\}_{p=1}^{\infty}$ *be as in* Lemma 5 *and let* K *be the subset of* $\{1, 2, \cdots, n\}$ *such that* $J \cap K = \phi$ *and* $J \cup K = \{1, 2, \cdots, n\}$*. Then, there exists a number* $\epsilon > 0$ *such that for all* $p \ge 1$ *and all* $k \in K$*,*

$$x_k(t) \ge \epsilon,\ \textit{for any}\ t \in [s_p, t_p].$$

Continuing the proof that Eqs.(9) and (15) $\Rightarrow$ Eq.(16), we note that if Eq.(16) does not hold and $x_i(t),\ 1 \le i \le n,\ J,\ and\ j_p$ are as in Lemma 5, there exists an integer $j_* \in J$ such that $j_p = j_*$ for infinitely many integers p. Let $\{p_q\}_{q=1}^{\infty}$ be an increasing sequence of integers such that

$$j_{p_q} = j_*,\ for\ any\ q \ge 1.$$

To simplify the notation, let $c_q = s_{p_q}$ and $d_q = t_{p_q}$ for $q \ge 1$, so

$$d_q - c_q \ge p_q,\ for\ any\ q \ge 1.$$

Since, according to Eq.(22),

$$\max\{\ x_j(t) \mid j \in J,\ c_q \le t \le d_q\} \le \delta/p_q,$$

and for $0 \le l \le m$,

$$\frac{1}{d_q - c_q}\int_{c_q}^{d_q} x_j(t-\tau_l)dt = \frac{1}{d_q - c_q}\int_{c_q}^{d_q} x_j(t)dt + \frac{1}{d_q - c_q}\int_{c_q-\tau_l}^{c_q} x_j(t)dt - \frac{1}{d_q - c_q}\int_{d_q-\tau_l}^{d_q} x_j(t)dt,$$

we have that for any $j \in J$,

$$\lim_{q\to\infty}\frac{1}{d_q - c_q}\int_{c_q}^{d_q} x_j(t-\tau_l)dt = 0,\ 0 \le l \le m.$$

Since, according to Eqs.(22) and (23), $x_{j_*}(c_q) = \delta/p_q \ge x_{j_*}(d_q)$, we have that for any $q \ge 1$,

$$\ln(x_{j_*}(d_q)/x_{j_*}(c_q)) \le 0. \tag{24}$$

The following lemma is an improved version of Lemma 2.6 in Ahmad and Lazer [2] (see Lemma 2.7 in Muroya [7]).

Lemma 7. *For Eqs.*(3) *and* (10)*, assume Eqs.*(11) *and* (15) *and suppose that Eq.*(16) *does not hold. Let J and K be as in Lemmas 5 and 6. Then there exists for each $k \in K$, a number $x_k > 0$ such that for any $k \in K$,*

$$M[c_k] \ge \bar{a}_{kL}x_k + \sum_{j\in K}\bar{b}_{kjL}x_j, \tag{25}$$

and there exists a $j_* \in J$ such that

$$m[c_{j_*}] \le \sum_{k\in K}\bar{b}_{j_*kM}x_k. \tag{26}$$

It is now easy to finish the proof that Eqs.(9) and (15) $\Rightarrow$ Eq.(16) in Theorem 1 by contradiction. Suppose that conditions (9) and (15) are true but Eq.(16) does not hold. Then, Eq.(9) implies Eq.(11) and by the above lemmas, there exist two subsets J and K of $\{1, 2, \cdots, n\}$, an integer $j_* \in J$ and positive numbers $x_k,\ k \in K$ such that Eqs.(25)

and (26) hold.
Let $x_j = 0$ for any $j \in J$ and $\boldsymbol{x} = (x_1, x_2, \cdots, x_n)^T$. From Eqs.(25) and (9), we see that

$$\begin{cases} M[c_k] \geq \bar{a}_{kL}x_k + \sum\limits_{j\in K} \bar{b}_{kjL}x_j \geq \bar{a}_{kL}x_k + \sum\limits_{j=1}^{n} \bar{b}^{-}_{kjL}x_j, \ \textit{for any } k \in K, \\ M[c_i] \geq 0 \geq \bar{a}_{iL}x_i + \sum\limits_{j=1}^{n} \bar{b}^{-}_{ijL}x_j, \ \textit{for any } i \in J. \end{cases}$$

Then, by Eqs.(1) and (9), it holds that

$$\boldsymbol{x} \leq (\bar{A}_L + \bar{B}^{-}_L)^{-1}\bar{\boldsymbol{c}}. \tag{27}$$

It follows from Eq.(26) that

$$m[c_{j_*}] \leq \sum_{k\in K} \bar{b}_{j_*kM}x_k \leq \sum_{k\in K} \bar{b}^{+}_{j_*kM}x_k.$$

However, by conditions (9) and (27), we see that

$$\underline{\boldsymbol{c}} > \bar{B}^{+}_M(\bar{A}_L + \bar{B}^{-}_L)^{-1}\bar{\boldsymbol{c}} \geq \bar{B}^{+}_M\boldsymbol{x}.$$

Therefore, by Eq.(6), it holds that

$$m[c_{j_*}] > \sum_{k\neq j_*} \bar{b}^{+}_{j_*kM}x_k = \sum_{k\in K} \bar{b}^{+}_{j_*kM}x_k.$$

This contradiction proves that Eqs.(9) and (15) $\Rightarrow$ Eq.(16). □

The following lemma is a basic result to obtain a sufficient conditions of the global asymptotic stability of system (1)-(2).

Lemma 8. *For any two solutions* $\{x_i(t)\}_{i=1}^n$ *and* $\{y_i(t)\}_{i=1}^n$ *of the system* (1)-(2)*, it holds Eq.*(13).

Proof. We have that

$$\begin{aligned} &(x_i(t) - x_i(t-\tau_l)) - (y_i(t) - y_i(t-\tau_l)) \\ &\qquad = \int_{t-\tau_l}^{t} x_i'(s)ds - \int_{t-\tau_l}^{t} y_i'(s)ds \\ &\qquad = \int_{t-\tau_l}^{t} \{c_i(s) - \sum_{j=1}^{n}\sum_{k=0}^{m} a_{ij}^k(s)y_j(s-\tau_k)\}(x_i(s) - y_i(s))ds \\ &\qquad\quad - \int_{t-\tau_l}^{t} \sum_{j=1}^{n}\sum_{k=0}^{m} x_i(s)a_{ij}^k(s)(x_j(s-\tau_k) - y_j(s-\tau_k))ds, \\ &\qquad\qquad\qquad\qquad\qquad\qquad t \geq t_0 + \tau_l, \ 1 \leq i \leq n, \ 0 \leq l \leq m. \end{aligned}$$

On the other hand, by Eq.(1), we have that

$$\begin{aligned} \frac{x_i'(t)}{x_i(t)} - \frac{y_i'(t)}{y_i(t)} &= -\sum_{j=1}^{n}\sum_{l=0}^{m} \bar{a}_{ij}^l(t)(x_j(t-\tau_l) - y_j(t-\tau_l)) \\ &\quad - \sum_{l=0}^{m} a_{ii}^l(t)[(x_i(t) - y_i(t)) - \{(x_i(t) - x_i(t-\tau_l)) - (y_i(t) - y_i(t-\tau_l))\}], \end{aligned}$$

from which we obtain Eq.(13). □

Now, consider the proof that Eqs.(9), (15) and (17) ⇒ Eq.(18) in Theorem 1. The proof relies on ideas already used by Gopalsamy [4], Tineo and Alvalez [10], Redheffer [8], Ahmad and Lazer [1], [2] and Muroya [7] (see also Teng and Chen [9]).

Lemma 9. *In addition to the conditions* (9), (15) *and* (17), *suppose that in the system* (13)-(14), *there exist positive constants* $\alpha_1, \alpha_2, \cdots, \alpha_n$ *and* $\eta > 0$ *such that*

$$\alpha_i \bar{\bar{a}}_{ii} - \sum_{j \neq i} \alpha_j \bar{\bar{a}}_{ji} \geq \eta, \ 1 \leq i \leq n. \tag{28}$$

Then any two solutions $x_i(t), \ y_i(t), \ 1 \leq i \leq n$ *of system* (1)-(2), *satisfy the conditions*

$$\lim_{t \to \infty} (x_i(t) - y_i(t)) = 0, \ 1 \leq i \leq n.$$

Proof. For the positive constants $\alpha_1, \alpha_2, \cdots, \alpha_n$ in Eq.(28), consider a continuous Lyapunov-like (non-negative and non-differentiable) function $v(t)$ such that for $t \geq t_0$,

$$\begin{aligned} v(t) = & \sum_{i=1}^{n} \alpha_i [|\ln(x_i(t)/y_i(t))| + \sum_{j=1}^{n} \sum_{l=0}^{m} \int_{t-\tau_l}^{t} |\bar{a}_{ij}^l(s+\tau_l)||x_j(s) - y_j(s)| ds \\ & + \sum_{l=0}^{m} \int_{t-\tau_l}^{t} a_{ii}^l(s+\tau_l) \int_s^t |c_i(u) - \sum_{j=1}^{n} \sum_{k=0}^{m} a_{ij}^k(u) y_j(u-\tau_k)||x_i(u) - y_i(u)| du ds \\ & + \sum_{l=0}^{m} \int_{t-\tau_l}^{t} a_{ii}^l(s+\tau_l) \int_s^t \sum_{j=1}^{n} \sum_{k=0}^{m} |x_i(u) a_{ij}^k(u)||x_j(u-\tau_k) - y_j(u-\tau_k)| du ds \\ & + \sum_{l=0}^{m} \sum_{j=1}^{n} \sum_{k=0}^{m} \int_{t-\tau_k}^{t} (\int_{s-\tau_l+\tau_k}^{s+\tau_k} a_{ii}^l(u+\tau_l) du) |x_i(s+\tau_k) a_{ij}^k(s+\tau_k)||x_j(s) - y_j(s)| ds]. \end{aligned}$$

One can verify that

$$x(t)\sigma(x)(t) = |x(t)| \ and \ D^+|x(t)| = \sigma(x)(t) \frac{dx(t)}{dt},$$

where

$$\sigma(x)(t) = \begin{cases} 1, & if \ x(t) > 0; or \ x(t) = 0 \ and \ \frac{dx(t)}{dt} > 0, \\ 0, & if \ x(t) = 0 \ and \ \frac{x(t)}{dt} = 0, \\ -1, & if \ x(t) < 0; or \ x(t) = 0 \ and \ \frac{x(t)}{dt} < 0 \end{cases}$$

and D^+ denotes the upper right Dini-derivative. Since the signs of $\ln(x_i(t)/y_i(t))$ and $x_i(t) - y_i(t)$ are same, it follows from Eq.(13) and the above that for $t \geq t_0$ *and* $1 \leq i \leq n$,

$$\begin{aligned} D^+|\ln(x_i(t)/y_i(t))| \leq & -\bar{a}_i(t)|x_i(t) - y_i(t)| + \sum_{j=1}^{n} \sum_{l=0}^{m} |\bar{a}_{ij}^l(t)||x_j(t-\tau_l) - y_j(t-\tau_l)| \\ & + \sum_{l=0}^{m} a_{ii}^l(t) \int_{t-\tau_l}^{t} |\{c_i(s) - \sum_{j=1}^{n} \sum_{k=0}^{m} a_{ij}^k(s) y_j(s-\tau_k)\}(x_i(s) - y_i(s)) \\ & - \sum_{j=1}^{n} \sum_{k=0}^{m} x_i(s) a_{ij}^k(s)(x_j(s-\tau_k) - y_j(s-\tau_k))| ds. \end{aligned} \tag{29}$$

By Eqs.(9) and (15) $\Rightarrow$ Eq.(16) in Theorem 1, $x_i(t)$ *and* $y_i(t),\ 1 \le i \le n$ are bounded above and below by positive constants for $t \ge t_0$. Therefore, it follows that for any $t \ge t_0 + \max_{0\le l\le m} \tau_l$, $v(t) < +\infty$.
Calculating the upper right Dini-derivative of v and using Lemma 2, Eqs.(28), (29) and (14), we obtain that there exists a sufficiently small $\epsilon > 0$ such that $0 < \eta/2 \le \eta_\epsilon$ and for a sufficiently large $t \ge t_0$, $x_i(t), y_i(t) < \bar{x}_i + \epsilon,\ 1 \le i \le n$ and

$$
\begin{aligned}
&D^+v(t)\\
&\le -\sum_{i=1}^{n} \alpha_i[\{\bar{a}_i(t) - \sum_{l=0}^{m}(\int_{t-\tau_l}^{t} a_{ii}^l(s+\tau_l)ds)|c_i(t) - \sum_{j=1}^{n}\sum_{k=0}^{m} a_{ij}^k(t)y_j(t-\tau_k)|\}|x_i(t) - y_i(t)|\\
&\quad -\sum_{j=1}^{n}\sum_{l=0}^{m}\{|\bar{a}_{ij}^l(t+\tau_l)| + \sum_{k=0}^{m}(\int_{t-\tau_l+\tau_k}^{t+\tau_k} a_{ii}^l(s+\tau_l)ds)|x_i(t+\tau_k)a_{ij}^k(t+\tau_k)|\}|x_j(t) - y_j(t)|]\\
&\le -\sum_{i=1}^{n}\{\alpha_i\bar{\tilde{a}}_{ii}^{\epsilon} - \sum_{j\ne i}\alpha_j\bar{\tilde{a}}_{ji}^{\epsilon}\}|x_i(t) - y_i(t)|\\
&\le -\frac{\eta}{2}\sum_{i=1}^{n}|x_i(t) - y_i(t)|,
\end{aligned}
$$

where

$$
\begin{cases}
\bar{\tilde{a}}_{ii}^{\epsilon} = \tilde{a}_{iL} - (\sum_{l=0}^{m} a_{iiM}^l\tau_l)\{\bar{a}_{iM}(\bar{x}_i+\epsilon)\\
\qquad + \max(-c_{iL} + (\bar{a}_{iM}(\bar{x}_i+\epsilon) + \sum_{j=1}^{n}\bar{b}_{ij}^{+}(\bar{x}_j+\epsilon)), c_{iM} - \sum_{j=1}^{n}\bar{b}_{ij}^{-}(\bar{x}_j+\epsilon))\},\\
\bar{\tilde{a}}_{ij}^{\epsilon} = (-b_{ijL}^{-} + b_{ijM}^{+})\{1 + (\sum_{l=0}^{m} a_{iiM}^l\tau_l)(\bar{x}_i+\epsilon)\},\ j \ne i,\ 1 \le i \le n,\\
\eta_\epsilon \equiv \min_{1\le i\le n}\{\alpha_i\bar{\tilde{a}}_{ii}^{\epsilon} - \sum_{j\ne i}\alpha_j\bar{\tilde{a}}_{ji}^{\epsilon}\} > 0.
\end{cases}
$$

Hence, the remained part of proof is similar to the proof in Gopalsamy [4] and Ahmad and Lazer [2]. Thus, we get the conclusion. □

We prove that the condition (17) implies Eq.(28) (see Lemma 2.9 in Muroya [7] and cf. Lemma 3.2 in Ahmad and Lazer [2]).

Lemma 10. (*See Berman and Plemmons [3]). The condition* (17) *holds, if and only if, there exist constants* $\alpha_i > 0,\ 1 \le i \le n$ *such that*

$$\alpha_i\bar{\tilde{a}}_{ii} - \sum_{j\ne i}\alpha_j\bar{\tilde{a}}_{ji} > 0,\ 1 \le i \le n. \tag{30}$$

Proof. Let $\bar{A} = \bar{\tilde{A}} = [\bar{\tilde{a}}_{ij}]$, $\bar{D} = \mathrm{diag}(\bar{\tilde{a}}_{11}, \bar{\tilde{a}}_{22}, \cdots, \bar{\tilde{a}}_{nn})$ and $\bar{B} = I - \bar{D}^{-1}\bar{A}$ be $n \times n$ matrices, where $\bar{\tilde{a}}_{ii} > 0,\ 1 \le i \le n$ and I is the $n \times n$ unit matrix. Then, it is well known (see for example, Berman and Plemmons [3]) that $\bar{A}$ is an M-matrix, if and only if,

$$\bar{\tilde{a}}_{ii} > 0,\ 1 \le i \le n,\ \bar{B} \ge \mathbf{0}\ \textit{and}\ \rho(\bar{B}) < 1,$$

where $\rho(\bar{B})$ is the spectral radius of matrix $\bar{B}$.
Then,

$$\begin{cases} \bar{A} = \bar{D}(I - \bar{B}),\ \bar{A}^T = (I - \bar{B}^T)\bar{D} = \bar{D}(I - \bar{D}^{-1}\bar{B}^T\bar{D}),\ \bar{D}^{-1}\bar{B}^T\bar{D} \geq \mathbf{0} \\ \textit{and } \rho(\bar{D}^{-1}\bar{B}^T\bar{D}) = \rho(\bar{B}^T) = \rho(\bar{B}) < 1, \end{cases}$$

which implies that $\bar{A}^T$ is also an M-matrix.
It is also well known (see Berman and Plemmons [3]) that $\bar{A}^T$ is an M-matrix, if and only if, there is a diagonal matrix $\tilde{D} = \mathrm{diag}(\alpha_1, \alpha_2, \cdots, \alpha_n)$ such that

$$\alpha_i > 0,\ 1 \leq i \leq n,\ \textit{and } \bar{A}^T\tilde{D} \textit{ is a diagonally dominant matrix},$$

which implies Eq.(30).
Thus, we get the conclusion. □

By Lemmas 9 and 10, we complete the proof that Eqs.(9), (16) and (17) ⇒ Eq.(18) in Theorem 1.

Now, consider the following Lotka-Volterra type competitive system with discrete delays (see Zhen and Ma [12]):

$$\begin{cases} x'(t) = x(t)\{c_1 - a_{11}x(t - \tau_{11}) - a_{12}y(t - \tau_{12})\}, \\ y'(t) = y(t)\{c_2 - a_{21}x(t - \tau_{21}) - a_{22}y(t - \tau_{22})\}, \end{cases} \tag{31}$$

with initial conditions

$$\begin{cases} x(t) = \phi_1(t) \geq 0,\ t \in [-\tau, 0],\ \phi_1(0) > 0, \\ y(t) = \phi_2(t) \geq 0,\ t \in [-\tau, 0],\ \phi_2(0) > 0, \\ \tau = \max\{\tau_{ij}\}, \end{cases}$$

where $x(t), y(t)$ stand for densities of both the population at time t, respectively, c_i, a_{ij} are all positive constants, and τ_{ij} are nonnegative.

To this system, in [12], Zhen and Ma have obtained the following result.

Theorem A. (*Zhen and Ma [12]*).
Suppose that

$$\frac{a_{11}}{a_{21}} > \frac{c_1}{c_2} > \frac{a_{12}}{a_{22}}, \tag{32}$$

and

$$a_{ii}\tau_{ii}M_i < 1,\ i = 1, 2,\ \textit{and } A_{11}A_{22} - A_{12}A_{21} > 0, \tag{33}$$

where

$$\begin{cases} A_{11} = a_{11}(1 - a_{11}\tau_{11}M_1),\ A_{12} = -a_{12}(1 + a_{11}\tau_{11}M_1), \\ A_{21} = -a_{21}(1 + a_{22}\tau_{22}M_2),\ A_{22} = a_{22}(1 - a_{22}\tau_{22}M_2), \\ M_1 = (\frac{c_1}{a_{11}})e^{c_1\tau_{11}} \textit{ and } M_2 = (\frac{c_2}{a_{22}})e^{c_2\tau_{22}}. \end{cases}$$

Then the positive equilibrium (x^*, y^*) *of system* (31) *is globally asymptotically stable.*

To the system (31), by (32) implies (15) for $Q = \{1, 2\}$ and the existence of the positive equilibrium (x^*, y^*) of system (31). Then, by the particular case in Theorem 1, we have that

$$\bar{\bar{a}}_{ii} = A_{ii},\ \bar{\bar{a}}_{ij} = A_{ij},\ j \neq i,\ i = 1, 2, \tag{34}$$

and (17) is equivalent to (33). Hence, by Theorem 1, we obtain the same conclusion of Theorem A. Thus, Theorem 1 is an extension of Theorem A to nonautonomous cases of several species.

Similarly, we consider the following time-dependent pure-delay-type Lotka-Volterra predator-prey model (see Xu and Chen [11]):

$$\begin{cases} \dfrac{dx_1(t)}{dt} = x_1(t)(c_1(t) - a_{11}(t)x_1(t-\tau_{11}) - a_{12}(t)x_2(t-\tau_{12})), \\ \dfrac{dx_2(t)}{dt} = x_2(t)(-c_2(t) + a_{21}(t)x_1(t-\tau_{21}) - a_{22}(t)x_2(t-\tau_{22}) - a_{23}(t)x_3(t-\tau_{23})), \\ \dfrac{dx_3(t)}{dt} = x_3(t)(-c_3(t) + a_{32}(t)x_2(t-\tau_{32}) - a_{33}(t)x_3(t-\tau_{33})), \end{cases} \tag{35}$$

where $c_i(t)$ and $a_{ij}(t),\ 1 \leq i, j \leq 3$ are continuous, bounded, and strictly positive functions on $[0, +\infty)$, and $\tau_{ij},\ 1 \leq i, j \leq 3$ are nonnegative constants.

To this system, in [11], Xu and Chen have obtained the following result.

Theorem B. (*see Xu and Chen [11]*).
Assume

$$\begin{cases} a_{32M}(a_{21M}c_{1M}\exp(\tau_{11}c_{1M}) - c_{2L}a_{11L})\exp\{(\frac{a_{21M}c_{1M}}{a_{11L}}\exp(\tau_{11}c_{1M}) - c_{2L})\tau_{22}\} \\ \quad -a_{11L}a_{22L}c_{3L} > 0, \\ a_{32L}(a_{21L}\underline{x}_1 - c_{2M} - a_{23M}\bar{x}_3)/a_{22M} \\ \qquad \times \exp\{(a_{21L}\underline{x}_1 - c_{2M} - a_{22M}\bar{x}_2 - a_{23M}\bar{x}_3)\tau_{22}\} - c_{3M} > 0, \end{cases} \tag{36}$$

where

$$\begin{cases} \bar{x}_1 = \frac{c_{1M}}{a_{11L}}\exp(\tau_{11}c_{1M}),\ \bar{x}_2 = \frac{a_{21M}\bar{x}_1 - c_{2L}}{a_{22L}}\exp\{(a_{21M}\bar{x}_1 - c_{2L})\tau_{22}\}, \\ \bar{x}_3 = \frac{a_{32M}\bar{x}_2 - c_{3L}}{a_{33L}}\exp\{(a_{32M}\bar{x}_2 - c_{3L})\tau_{33}\}, \\ \underline{x}_1 = \frac{c_{1L} - a_{12M}\bar{x}_2}{a_{11M}}\exp\{(c_{1L} - a_{11M}\bar{x}_1 - a_{12M}\bar{x}_2)\tau_{11}\}, \\ \underline{x}_2 = \frac{a_{21L}\underline{x}_1 - c_{2M} - a_{23M}\bar{x}_3}{a_{22M}}\exp\{(a_{21L}\underline{x}_1 - c_{2M} - a_{22M}\bar{x}_2 - a_{23M}\bar{x}_3)\tau_{22}\}, \\ \underline{x}_3 = \frac{a_{32L}\underline{x}_2 - c_{3M}}{a_{33M}}\exp\{(a_{32L}\underline{x}_2 - c_{3M} - a_{33M}\bar{x}_3)\tau_{33}\}, \end{cases}$$

and suppose further that the system (35) *satisfies*

$$A_i > 0,\ i = 1, 2, 3, \tag{37}$$

where

$$\begin{cases} b_{iiL} = a_{iiL}\underline{x}_i,\ b_{ijM} = a_{ijM}\bar{x}_i,\ i = 1, 2, 3, \\ A_1 = 2b_{11L} - b_{12M} - b_{21M} - b_{11M}(2b_{11M} + b_{12M})\tau_{11} - b_{21M}b_{22M}\tau_{22}, \\ A_2 = 2b_{22L} - b_{12M} - b_{21M} - b_{23M} - b_{32M} - b_{11M}b_{12M}\tau_{11} \\ \qquad -b_{22M}(b_{21M} + 2b_{22M} + b_{23M})\tau_{22} - b_{32M}b_{33M}\tau_{33}, \\ A_3 = 2b_{33L} - b_{23M} - b_{32M} - b_{22M}b_{23M}\tau_{22} - b_{33M}(b_{32M} + 2b_{33M})\tau_{33}. \end{cases}$$

Then any positive solution of system (35) *is globally asymptotically stable.*

To the system (35), we apply Theorem 1, then we obtain the following result.

Corollary 1. *Assume* (36) *and suppose that*

$$\begin{cases} \bar{\bar{a}}_{11}, \bar{\bar{a}}_{22}, \bar{\bar{a}}_{33} > 0, \ \textit{and} \\ \bar{\bar{a}}_{12}\bar{\bar{a}}_{21}\bar{\bar{a}}_{33} + \bar{\bar{a}}_{23}\bar{\bar{a}}_{32}\bar{\bar{a}}_{11} < \bar{\bar{a}}_{11}\bar{\bar{a}}_{22}\bar{\bar{a}}_{33}, \end{cases} \tag{38}$$

where

$$\begin{cases} \bar{x}_1 = \frac{c_{1M}}{a_{11L}} \exp(\tau_{11}c_{1M}), \ \bar{x}_2 = \frac{a_{21M}\bar{x}_1 - c_{2L}}{a_{22L}} \exp\{(a_{21M}\bar{x}_1 - c_{2L})\tau_{22}\}, \\ \bar{x}_3 = \frac{a_{32M}\bar{x}_2 - c_{3L}}{a_{33L}} \exp\{(a_{32M}\bar{x}_2 - c_{3L})\tau_{33}\}, \end{cases} \tag{39}$$

and

$$\begin{cases} \bar{\bar{a}}_{11} = a_{11L} - (a_{11M}\tau_{11})\{a_{11M}\bar{x}_1 + \max(c_{1M}, -c_{1L} + a_{12}\bar{x}_2)\}, \\ \bar{\bar{a}}_{22} = a_{22L} - (a_{22M}\tau_{22})\{a_{22M}\bar{x}_2 + \max(-c_{2L} + a_{21M}\bar{x}_1, c_{2M} + a_{23}\bar{x}_3)\}, \\ \bar{\bar{a}}_{33} = a_{33L} - (a_{33M}\tau_{33})\{a_{33M}\bar{x}_3 + \max(-c_{3L} + a_{32M}\bar{x}_2, c_{3M})\}, \end{cases} \tag{40}$$

and

$$\begin{cases} \bar{\bar{a}}_{12} = a_{12M}(1 + a_{11M}\bar{x}_1), \ \bar{\bar{a}}_{21} = a_{21M}(1 + a_{22M}\bar{x}_2), \\ \bar{\bar{a}}_{23} = a_{23M}(1 + a_{22M}\bar{x}_2), \ \bar{\bar{a}}_{32} = a_{32M}(1 + a_{33M}\bar{x}_3). \end{cases} \tag{41}$$

Then any positive solution of system (35) *is globally asymptotically stable.*

Proof. Eq.(36) implies the permanence of system (35) and hence Eq.(16) holds (see Xu and Chen [11]). Since 3×3 tridiagonal matrix $\bar{\bar{A}} = [\bar{\bar{a}}_{ij}]$ is an M-matrix, if and only if, $\bar{\bar{a}}_{11}, \bar{\bar{a}}_{22}, \bar{\bar{a}}_{33} > 0 \geq \bar{\bar{a}}_{12}, \bar{\bar{a}}_{21}, \bar{\bar{a}}_{23}, \bar{\bar{a}}_{32}$ and $(\bar{\bar{a}}_{12}/\bar{\bar{a}}_{11})(\bar{\bar{a}}_{21}/\bar{\bar{a}}_{22}) + (\bar{\bar{a}}_{23}/\bar{\bar{a}}_{22})(\bar{\bar{a}}_{32}/\bar{\bar{a}}_{33}) < 1$, for this system (35), we can easily prove that Eq.(17) holds if and only if Eq.(38) holds (see for example, Berman and Plemmons [3]). Thus, by Theorem 1, we obtain the conclusion. □

Note that Eqs.(37) and (38) are different type conditions from each other.

Acknowledgement. Research partially supported by Waseda University Grant for Special Research Projects 2001A-571 and 2003A-573.

References

[1] S. Ahmad and A. C. Lazer, Necessary and sufficient average growth in a Lotka-Volterra system, *Nonlinear Analysis TMA* **34**, 191-228 (1998).

[2] S. Ahmad and A. C. Lazer, Average conditions for global asymptotic stability in a nonautonomous Lotka-Volterra system, *Nonlinear Analysis TMA* **40**, 37-49 (2000).

[3] A. Berman and R. J. Plemmons, *Nonnegative Matrices in the Mathematical Sciences*, Academic Press, New York, 1979.

[4] K. Gopalsamy, Global asymptotic stability in Volterra's population systems, *J. Math. Biol.* **19**, 157-168 (1984).

[5] K. Gopalsamy, Global asymptotic stability in a periodic Lotka-Volterra system, *J. Austral. Math. Soc. Ser B* **27**, 66-72 (1985).

[6] K. Gopalsamy, Global asymptotic stability in an almost-periodic Lotka-Volterra system, *J. Austral. Math. Soc. Ser B* **27**, 346-360 (1986).

[7] Y. Muroya, Persistence and global stability for nonautonomous Lotka-Volterra delay differential systems, *Communications on Applied Nonlinear Analysis* **9**, 31-45 (2002).

[8] R. Redheffer, Nonautonomous Lotka-Volterra system I, *J. Differential Equations* **127**, 519-540 (1996).

[9] Z. Teng and L. Chen, Global asymptotic stability of periodic Lotka-Volterra systems with delays, *Nonlinear Analysis* **45**, 1081-1095 (2001).

[10] A. Tineo and C. Alvarez, A different consideration about the globally asymptotically stable solution of the periodic n-competing species problem, *J. Math. Anal. Appl.* **159**, 44-60 (1991).

[11] R. Xu and L. Chen, Persistence and global stability for a delayed nonautonomous predator-prey system without dominating instantaneous negative feedback, *J. Math. Anal. Appl.* **262**, 50-61 (2001).

[12] J. Zhen and Z. Ma, Stability for a competitive Lotka-Volterra system with delays, *Nonlinear Analysis TMA* **51**, 1131-1142 (2002).

JOURNAL OF CONCRETE AND APPLICABLE MATHEMATICS,VOL.3,NO.3,283-311,2005,COPYRIGHT 2005 EUDOXUS PRESS,LLC

A General Theory of Stochastic Roundoff Error Analysis with Applications to DFT and DCT

Hansmartin Zeuner
PAV Card
hmzeuner@gmx.de

Abstract. In the first part of this article, a stochastic model for the average case analysis of roundoff errors is developed. Starting from the fractional roundoff error of an algebraic operation a Wilkinson type model $\mathrm{fl}(X \circ Y) = X \circ Y + \varepsilon^{\circ}|X \circ Y|$ for the relative roundoff error ε° is derived for real valued operations in §1 and for complex operations in §2. The analysis of alternative implementations of complex multiplication shows that the standard method has roundoff errors of the smallest average size. In §3 this model is applied to the error analysis of the matrix–vector product. The average Euclidean norm of the roundoff error vector as well as the covariance between its components is determined. In the special case of Givens rotations these components are uncorrelated and the same holds for reflections. For rotations, the standard method of the matrix–vector product is compared with the representation by three lifting steps which enjoys roundoff errors of smaller average size if the rotation angle is small, however at the expense of a correlation between the components.

In the last part, these results are applied to the discrete Fourier transform and the discrete cosine transforms of types II and III in the case $n = 8$. Taking into account the precomputation errors of the coefficients of the trigonometric matrices, both the direct method and a number of fast algorithms are studied. The latter show roundoff errors of only slightly larger average norm. In all algorithms the components of the roundoff errors are either uncorrelated or have covariances of only small size. The results are confirmed by simulations.

0. Introduction

When an algorithm which uses a large number of floating point operations is implemented, it is important to analyse the size of the roundoff errors in order to ensure the accuracy of the result. By a *worst case* analysis an upper bound of the error is obtained. Usually, this bound is much larger than the "typical" roundoff error of the algorithm: whereas the worst case error grows like the number of steps in the algorithm, in simulations often an error is observed which is proportional to the square root thereof (see [7], pp. 52–53). When we have to choose between different algorithms for the same problem, then in the case of equal complexity the *average* size of the errors would be more conclusive than a theoretical upper bound.

The *average case* analysis attempts to obtain a more realistic estimate for the roundoff errors by applying a stochastic model for these errors. Certainly, roundoff errors are *not random,* since the computer will produce the same result again, and hence the same error, when an operation is repeated with the same operands. But, as in other areas of applied mathematics, we develop a *model* of the roundoff errors which well describes the effects observed in the data. It will turn out at the end

1991 *Mathematics Subject Classification.* Primary 65G05, 65T50.

Key words and phrases. Roundoff error analysis, discrete Fourier transform, fast Fourier transform, discrete cosine transform.

that the predictions of our stochastic model for the roundoff errors are in good accordance with the results of simulations.

In the early days of the computer, average case analysis of roundoff errors has been mainly used to evaluate the benefits of different hardware concepts and implementations of the basic floating point arithmetic (see [3]). For special distributions of the data, the statistical properties of the roundoff errors in arithmetic expressions have been studied. Applications to the a posteriori average case analysis of the fast Fourier transform (FFT) are proved in [4] and [1] for the Cooley–Tukey resp. Gentleman–Sande algorithms – under the assumption that the roots of unity used in the FFT have been precomputed exactly. It turns out that these precomputation errors have a determinative influence on the accuracy of the result (see [11]). In [13], an average case error analysis for the Cooley–Tukey FFT is given which includes the treatment of the precomputation errors.

In the first half of this article we lay the ground for a simple but realistic model for the a priori average case analysis of the roundoff errors in floating point arithmetic. It is valid both for real and complex valued data. In Section 1 we derive a Wilkinson–type (but stochastic) model for the *relative error* in the case of *real valued* arithmetic, which is based on properties of the *fractional error* as defined in [3]. For multiplication (and division) the distribution of the relative error can be given in closed form. The dependency from the distribution of the data is only very small. In contrast, for summation, the distribution of the data has a strong influence on the distribution of the relative error. A formula for the expectation of the relative error is proved under various rounding modes, and it is shown that the expectation is zero for the default rounding mode in the IEEE 754 standard.

In Section 2 properties of the relative roundoff error $\varepsilon^\circ := \frac{\mathrm{fl}(Z \circ W) - Z \circ W}{|Z \circ W|}$ for *complex* operands Z, W are derived from the stochastic Wilkinson model in Section 1. For all operations $\circ$, the real and the imaginary parts of the error are uncorrelated and have symmetric distributions with equal variances. For addition and multiplication the variance of the complex error does not depend on the distribution of the data, but only from the first two moments of the real valued relative errors. In the case of division there is only a weak dependence on the distribution of the data. Alternative methods for complex multiplication are shown to have inferior accuracy.

In the second half of the article the stochastic model for the actual roundoff error is applied to matrix–vector multiplication and the discrete Fourier and cosine transforms (DFT and DCT) in particular. In Section 3 we state our final model which is based on the preliminary models of Sections 1–2. The main assertion of the model is that the relative roundoff error ε° has expectation 0 and variance $\sigma^2_{\mathbb{K},\circ} u^2$ and is independent of the data. The constant $\sigma^2_{\mathbb{K},\circ}$ only depends on the operation $\circ$ and the underlying number field $\mathbb{K} = \mathbb{R}$ or $\mathbb{C}$ but not on the machine accuracy u (see (1.1)). Motivated by the results of section 2 we also assume that $\sigma^2_{\mathbb{K},\circ}$ is *not* dependent on the distribution of X and Y, as long as the data has expectation 0 and is uncorrelated. Based on this model, we determine the variance of the roundoff error for the inner product of two vectors and the product of a matrix with a vector under different summation methods. It is possible to generalize this to data with non–zero expectation, but the statement of the result is more complicated.

If entries of the matrix carry precomputation errors, several additional problems have to be considered: in many practical examples, the same constant occurs in

several entries of the matrix. The corresponding roundoff errors are clearly dependent. Moreover, multiplications with identical operands – and hence identical errors – enter in different components of the result, for example in the butterfly operations of the FFT. As a consequence, the components of the total roundoff error of a matrix–vector product are no longer independent. A general formula for the covariance matrix of the error vector in this situation is the main result of Section 3. If we are only interested in the mean square norm of the error vector (and not in the componentwise analysis), the result can be simplified considerably.

In the Section 4 we apply this general result to the discrete Fourier transform, the fast Cooley–Tukey and Gentleman–Sande algorithms, the discrete cosine–II and –III transforms with two new fast algorithms from recent research by Loeffler et al. [9] and Schreiber [12]. We restrict the study to the special case $n = 8$, since here the effect of correlation between error components and non–equal variances is more pronounced than for larger n. In all these examples, we observe that the correlations between the components of the error vector – if non–zero – are always small compared to the variances, but that the size of the variance of different components can vary considerably. The theoretical results are confirmed by simulations.

1. Wilkinson's roundoff error model revisited

The standard representation for roundoff errors occuring in computations with floating point numbers is Wilkinson's model, which gives the following estimate of the computed value of an operation $\circ \in \{+, -, \times, \div\}$

$$\mathrm{fl}(x \circ y) = (x \circ y)(1 + \varepsilon^\circ) \qquad (x, y \in \mathbb{R}). \tag{1.1}$$

The relative error ε° satisfies the worst case bound $|\varepsilon^\circ| \le u$, where u is the maximal relative machine error of the implementation of the arithmetic operations (depending on the CPU). If the mantissa in the floating point representation of a real number uses $m+1$ bits (not including the sign), then $u = 2^{-m-1}$. Here and in the following we will neglect the possibility of an underflow or overflow.

This model is based on the assumption that the computed result of an operation is obtained by rounding the precise result $z = x \circ y$ (or at least a more precise result) towards the nearest machine number

$$\mathrm{fl}(z) = z(1 + \varepsilon) \qquad (|\varepsilon| \le u).$$

A sharper and more realistic (but more complicated) model which is based on the representation of floating point numbers, follows from the assumption

$$\mathrm{fl}(z) = z + \zeta(z)\eta \qquad (|\eta| \le u)$$

with

$$\zeta(z) := \begin{cases} \operatorname{sgn}(z) \max\{2^n \le |z| : n \in \mathbb{Z}\} & \text{for } z \neq 0, \\ 0 & \text{for } z = 0. \end{cases}$$

Here, η is the *fractional error* (see [3]). It implies

$$\mathrm{fl}(x \circ y) = x \circ y + \zeta(x \circ y)\, \eta^\circ \qquad (x, y \in \mathbb{R},\ |\eta^\circ| \le u) \tag{1.2}$$

The difference between the implications of these models, however, is only small (see Remark 2.2). In the rest of the Section 1 our results will be based on the model (1.2) and we will study the consequences for (1.1).

Following Calvetti [4] we begin to build a stochastic model for the roundoff errors by assuming that the relative errors ε° in (1.1) from an operation $\circ$ are random with expectation

$$\mathbb{E}(\varepsilon^\circ) = \mu_\circ u \tag{1.3}$$

and variance

$$\mathbb{V}(\varepsilon^\circ) = \sigma_\circ^2 u^2. \tag{1.4}$$

The constants $\mu_\circ, \sigma_\circ \in \mathbb{R}$ will depend on the operation $\circ \in \{+, -, \times, \div\}$ and the distribution of the data. In practice the result of a particular computation, and hence the roundoff error, is (hopefully) *not* random. Therefore we explain the randomness of ε° as a consequence of the randomness of the input data. In this model we will assume that the input is random and that the relative error ε° is independent of the data. Our simulations are in accordance with relative error and data being uncorrelated, but since the roundoff error is a function of the data, our stronger assumption of independence is certainly a simplification. It is motivated by the observation that the size of the *relative* error ε° is not related to the size of the data, and that the last digits of the data, which have the strongest influence on the roundoff error, are approximately independent of the rest of the mantissa.

Calvetti [4] has observed that in numerical experiments $\mu_+ < 0$ is obtained for the addition, whereas $\mu_\times$ is close to zero. In her simulations the roundoff error is determined by rounding the exact result to a given precision. In our own experiments (using MATLAB on a Macintosh computer and an SGI workstation, and verifying $2 \cdot 10^6$ calculated values using long arithmetic) no statistically significant evidence against the hypotheses $\mu^+ = 0$ and $\mu^\times = 0$ was found (see Remark 1.8). It is clear that the implementation of the basic arithmetic operations in different hardware or software will influence the distributions of the roundoff errors. It is therefore possible that the "drift" towards zero for the addition observed by Calvetti originates from the treatment of *ties* in her model, i.e., a result $x \circ y$ which is exactly the mean between two consecutive machine numbers. In her experiments these are rounded towards the smaller neighbour by the usual `round` function. Following the IEEE 754 standard (see [7], pp. 46–48), an arithmetic operation is performed to arbitrary precision and then rounded towards one of the adjacent machine numbers using one of four methods. In the default rounding method, ties are rounded towards the nearest *even* mantissa number. Lemma 1.3 below will explain the discrepancy between Calvetti's model and ours.

In the following we will study the probability p° of ties for addition and multiplication and its influence on the expectation $\mu^\circ u$ of the roundoff error. We denote by $m + 1$ the number of bits in the mantissa (not counting the sign bit). The probability p^+ of a tie in the computation of a sum (or difference) depends strongly on the distribution of the operands.

1.1 Lemma. *Let X and Y be independent with $2^{j-1} \le X < 2^j$ and $2^{k-1} \le Y < 2^k$ $(j, k \in \mathbb{Z})$. Assume that the $|j-k|+1$ last digits in the binary representations are independent of the rest and uniformly distributed. Then the probability of a tie in the calculation of $X + Y$ is*

$$p^+ = 2^{-|j-k|-1}\big(1 + P\{X + Y < 2^{\max(j,k)}\}\big).$$

The calculation of $X-Y$ yields no roundoff error if $j=k$, and the probability of a tie is

$$p^+ = \begin{cases} \frac{1}{2}P\{|X-Y| \geq 2^{\max(j,k)-1}\} & \text{for } j = k \pm 1, \\ 2^{-|j-k|}\big(1 + P\{|X-Y| < 2^{\max(j,k)-1}\}\big) & \text{for } |j-k| \geq 2. \end{cases}$$

From these special cases the probability p^+ of a tie for arbitrary distributions of X and Y can be deduced. We observe that the probability of a tie can be as large as 50%, but very small probabilities are possible if X and Y are of very different size. Note that the assumption of independence is at least approximately true as long as $|j-k|$ is much smaller than m. Furthermore, the different behaviour of addition and subtraction stems from the assumption $X, Y > 0$.

Proof. Without loss of generality we may assume that $k \leq j = 0$. Let X_l and Y_l be the l-th digits in the binary representations of the mantissae of X and Y. Then we have a tie in the computation of $X+Y$ if and only if $X+Y \geq 1$ and $X_{m+1} + Y_{m+1+k} = 1, Y_{m+2+k} = \cdots = Y_{m+1} = 0$, or if $X+Y < 1$ and $Y_{m+2+k} = 1, Y_{m+3+k}, \ldots, Y_{m+1} = 0$. Since we have assumed independence, the probability of the first event is $P\{X+Y \geq 1\}\, 2^{k-1}$ and $P\{X+Y<1\}\, 2^k$ of the second. Note that $X+Y<1$ is only possible for $k \leq -1$. The sum of these probabilities is $2^{k-1}(1+P\{X+Y<1\})$, and so the first equation follows.

For $k \leq -1$ a tie in the difference $X-Y$ occurs if and only if $|X-Y| \geq \frac{1}{2}$ and $Y_{m+2+k} = 1, Y_{m+3+k}, \ldots, Y_{m+1} = 0$, or if $k \leq -2$, $|X-Y| < \frac{1}{2}$ and $Y_{m+3+k} = 1, Y_{m+4+k}, \ldots, Y_{m+1} = 0$. This implies the last two equations. ∎

1.2 Lemma. *If the density of the mantissa of X resp. Y (i.e. the probability that the mantissa is a particular number) is bounded by $a/2^{m+1}$ resp. $b/2^{m+1}$, then for the multiplication the probability $p^\times$ of a tie satisfies*

$$p^\times \leq \frac{3ab}{8} m 2^{-m}.$$

For example we can choose $a := \|f_X\|_\infty$ and $b := \|f_Y\|_\infty$ if X and Y are the machine approximations of random variables with densities f_X resp. f_Y and values in $[\frac{1}{2}, 1]$. In particular if X and Y are uniformly distributed in $[\frac{1}{2}, 1]$, $[0,1]$ or $[-1,1]$, then $a = b = 2$.

Proof. Without loss of generality let $\frac{1}{2} \leq X, Y < 1$. Let L resp. M be the number of trailing zeros in the mantissa of X resp. Y. Then

$$P\{L=k\} = \sum_{j=2^{m-k-1}}^{2^{m-k}-1} P\{X = \frac{2j+1}{2^{m-k+1}}\} \leq 2^{m-k-1} a/2^{m+1} = 2^{-k-2}a$$

and $P\{M=k\} \leq 2^{-k-2}b$ for $k = 0, \ldots, m-1$. A tie occurs if and only if $L+M = m-1$ and $XY \geq \frac{1}{2}$, or if $L+M = m-2$ and $XY < \frac{1}{2}$ (after normalization). Therefore the probability of a tie is bounded by

$$\begin{aligned} &P\{m-2 \leq L+M \leq m-1\} \\ &\quad = \sum_{k=0}^{m-1} P\{L=k\} P\{m-k-2 \leq M \leq m-k-1\} \\ &\quad \leq \sum_{k=0}^{m-1} 2^{-k-2} a (2^{-m+k} + 2^{-m+k-1}) b \quad = \quad 3mab2^{-m-3}. \end{aligned}$$

This completes the proof. ■

1.3 Lemma. *Let* $\circ \in \{+, -, \times, \div\}$ *and* $q := \mathbb{E}(\zeta(X \circ Y)/(X \circ Y))$ (*see* (1.2)). *Then*

$$\mu_\circ = \begin{cases} 0 & \textit{for nearest neighbour rounding with ties rounded to even,} \\ -p^\circ q & \textit{for nearest neighbour rounding with ties truncated,} \\ -q & \textit{for truncation.} \end{cases}$$

The constant q depends on the distribution of the data, but only slightly: if $X \circ Y$ is standard normal, then $q = (2\pi)^{-1/2} \sum_{n\in\mathbb{Z}} 2^{-n}(\exp(-2^{2n-1}) - \exp(-2^{2n+1})) \approx 0.721$; if $X \circ Y$ is uniform on $[-1, 1]$, then $q = \ln 2 \approx 0.693$.

Proof. Let $Z := X \circ Y$ and $R := Z/2\zeta(Z) - u\lfloor Z/2u\zeta(Z)\rfloor$ the part of the mantissa which would be lost with truncation. Since $0 \le R < u$, the distribution of R is symmetric around $u/2$ (with the exception of $R = 0$) and hence the first result follows. For truncation we have $\varepsilon^\circ = -2R\,\zeta(Z)/Z$ and $\mu_\circ u = \mathbb{E}(\varepsilon^\circ) = -2\mathbb{E}(R)\mathbb{E}(\zeta(Z)/Z) = -uq$ by the assumption of independence between relative error and data. For the second rounding method we have to replace $\mathbb{E}(R)$ by $\mathbb{E}(R\,1_{\{R=u/2\}}) = \frac{u}{2}P\{R = \frac{u}{2}\} = p^\circ\frac{u}{2}$. ■

In the following we will assume that the default rounding method of IEEE 754 standard is used. Hence $\mu_\circ = 0$. For summation the error distribution depends strongly on the distribution of the data (see Lemma 1.1 and the following histogramms of roundoff errors in Figures 1–2)

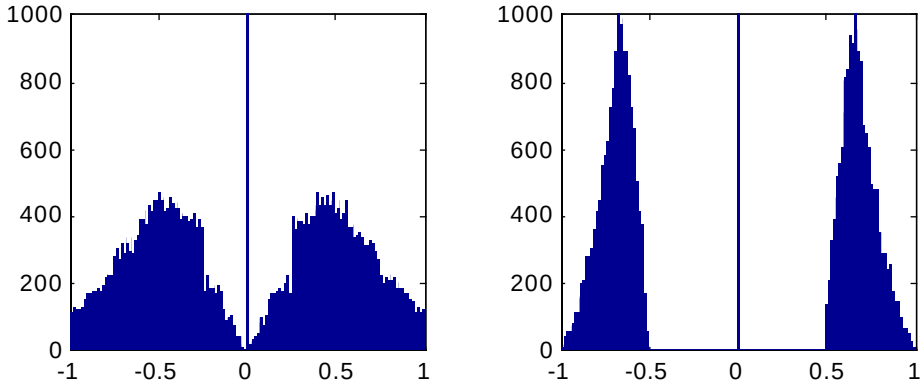


Figure 1: Distribution of roundoff errors ε^+/u from summation of standard Gaussian data (left) and data distributed uniformly on $[1, 2]$ (right), with 100000 samples.

In contrast, one observes that the distribution of the relative error for the multiplication has a very similar shape for a large class of input distributions. Simulations show that this shape can very well be described by a density:

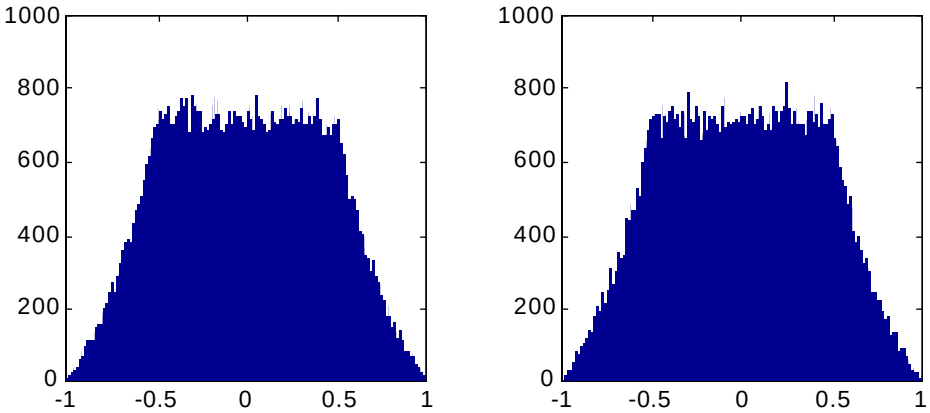


Figure 2: Distribution of roundoff errors $\varepsilon^{\times}/u$ from multiplication of standard Gaussian data (left) and data distributed uniformly on $[1,2]$ (right), with 100000 samples.

However, roundoff errors according to our definition are always discrete random variables.

In order to avoid this problem, we assume that the input variables are continuous with a smooth density, and rounded to machine number *after* multiplication. Then the roundoff error $\mathrm{fl}(XY) - XY$ also has a density. This modification is motivated by the fact that for uniformly distributed X and Y the remainder R (see Lemma 1.3) takes the value $uj/2^l$ with the same probability for all odd $j = 1, \dots, 2^l - 1$ ($l = 2, \dots, m+1$). Therefore the uniform distribution is a good approximation for the distribution of $\mathrm{fl}(XY) - XY$.

1.4 Proposition. *Assume that X and Y are independent with $\frac{1}{2} \le X, Y < 1$ and that X has a differentiable density ϕ with $u\|\phi'\|_\infty \ll 1$. Furthermore, let $\psi(t) := \mathbb{E}\big(Y\, 1_{\{Y \le t\}}\big)$ be the truncated moment. Then the relative error $\varepsilon^{\times}/u$ in (1.1) has the density θ on $[-1,1]$ given by*

$$\theta(t) \;=\; \mathbb{E}(X)\mathbb{E}(Y) + \int_{1/2}^{1} \psi\Big(\frac{1}{2x}\Big) x\phi(x)\, dx$$

for $|t| \le \frac{1}{2}$ and

$$\begin{aligned}\theta(t) \;&=\; \mathbb{E}(X\, 1_{\{2|t|X \le 1\}})\mathbb{E}(Y) - \int_{1/2}^{1} \psi\Big(\frac{1}{2x}\Big) x\phi(x)\, dx \\ &+\; 2\int_{1/2}^{1/2|t|} \psi\Big(\frac{1}{4x|t|}\Big) x\phi(x)\, dx + \int_{1/2|t|}^{1} \psi\Big(\frac{1}{2x|t|}\Big) x\phi(x)\, dx\end{aligned}$$

for $|t| \ge \frac{1}{2}$.

The assumption $u\|\phi'\|_\infty \ll 1$ means that the exact density of $\varepsilon^{\times}/u$ can be approximated by θ with a $\mathcal{O}(u\|\phi'\|_\infty)$–error term.

Proof. Under our assumptions the conditional distribution of $\mathrm{fl}(XY) - XY$ given X, Y is the uniform distribution on $[-\frac{u}{2}, \frac{u}{2}]$ for $XY \ge \frac{1}{2}$ and $[-\frac{u}{4}, \frac{u}{4}]$ for $XY < \frac{1}{2}$. Hence $\varepsilon^{\times}/u$ has the conditional density $XY\, 1_{[-\frac{1}{2XY}, \frac{1}{2XY}]}$ for $XY \ge \frac{1}{2}$ and $2XY\, 1_{[-\frac{1}{4XY}, \frac{1}{4XY}]}$ for $XY < \frac{1}{2}$. This holds up to an error of at most $u\|\phi'\|_\infty$.

Therefore the value $\theta(t)$ of the density of $\varepsilon^\times/u$ is

$$\iint_{xy<\frac12} 2xy\,1_{[0,\frac{1}{4|t|}]}(xy)P_Y(dy)\phi(x)dx + \iint_{xy\ge\frac12} xy\,1_{[0,\frac{1}{2|t|}]}(xy)P_Y(dy)\phi(x)dx$$
$$= 2\int_{1/2}^1 \psi\Big(\min\Big(\frac{1}{2x},\frac{1}{4x|t|}\Big)\Big)x\phi(x)dx + \int_{1/2}^1\Big[\psi\Big(\frac{1}{2x|t|}\Big)-\psi\Big(\frac{1}{2x}\Big)\Big]x\phi(x)dx.$$

For $|t|\le\frac12$ the value of the minimum in the integral above is $\frac{1}{2x}$ and $\psi(\frac{1}{2x|t|}) = \psi(1) = \mathbb{E}(Y)$. For $|t|\ge\frac12$ the minimum is $\frac{1}{4x|t|}$ and for $2x|t|\le 1$ we have $\psi(\frac{1}{2x|t|}) = \mathbb{E}(Y)$. In both cases the result follows by a simple calculation. ∎

1.5 Corollary. *Under the same assumptions as in Proposition* 1.4 *let* $\chi(t) := \mathbb{E}\big(Y^{-2}\,1_{\{Y\le t\}}\big)$. *Then the variance of the relative error* $\varepsilon^\times$ *in* (1.1) *is given by*

$$\sigma_\times^2 \;=\; \frac{1}{12}\,\mathbb{E}(X^{-2})\mathbb{E}(Y^{-2}) - \frac{1}{16}\int_{1/2}^1 \chi\Big(\frac{1}{2x}\Big)x\phi(x)\,dx.$$

1.6 Example. If X, Y are uniformly distributed on $[1/2, 1]$ (or $[0,1]$ or $[-1,1]$), then the density of $\varepsilon^\times/u$ is equal to

$$\theta(t) = \begin{cases} \frac38 + \frac12\ln 2 \;\approx\; 0.722 & \text{for } |t|\le\frac12\,,\\ \frac{\ln|t|}{4t^2} + \Big(\frac{\ln 2}{2}+\frac18\Big)\Big(\frac{1}{t^2}-1\Big) & \text{for } \frac12\le|t|\le 1.\end{cases}$$

The standard deviation is $\sigma_\times = \sqrt{\frac12\ln 2 - \frac16} \approx 0.424$.

1.7 Remark. For the division of real numbers, Proposition 1.4, Corollary 1.5 and the density in Example 1.6 are also valid, since the same arguments are applicable.

1.8 Remark. By experimenting with standard normal data X, Y in MATLAB we have obtained the following values as multiples of the machine accuracy u (which is $u = \frac12\texttt{eps} = 2^{-53}$ for MATLAB): $|\mu_+|, |\mu_\times|, |\mu_\div| < 0.001$, $\sigma_+ \approx 0.406$ and $\sigma_\times \approx 0.425 \approx \sigma_\div$. The results in [4] about μ_+ and $\mu_\times$ are explained by Lemmas 1.1–1.3.

For uniform data on $[0,1]$, $\sigma_+ \approx 0.470$ was obtained and $\sigma_+ \approx 0.192$ for the uniform distribution of $[-1,1]$. The influence of the distribution on the roundoff error from summation is obvious. For multiplication no such effect has been observed.

1.9 Remark. The subsequent results remain valid even when $\mu_+ \ne 0$ or $\mu_\times \ne 0$. However, if we replace the definition of the relative error in (1.1) by

$$\mathrm{fl}(X\circ Y) = X\circ Y + \varepsilon^\circ\,|X\circ Y|, \tag{1.5}$$

then it is easy to see that $\mathbb{E}(\varepsilon^\circ) = 0$ and $\mathbb{V}(\varepsilon^\circ) = \sigma_\circ^2 + \mu_\circ^2$ for any operation $\circ \in \{+,-,\times,\div\}$, as long as $|X\circ Y|$ and $\mathrm{sgn}(X\circ Y)$ are independent and $\mathrm{fl}(-x) = -\,\mathrm{fl}(x)$ for all $x\in\mathbb{R}$.

2. Complex arithmetic

Starting now from the stochastic version of Wilkinson's model (1.1) for real valued arithmetic operations, we will derive properties of the distributions of the relative roundoff errors for operations with complex data. These will be stated in the form of (1.5) which is more convenient for later use.

2.1 Proposition. *Let X, Y, U, V be independent real–valued random variables with identical distribution symmetric around the origin. Set $Z := X + \mathrm{i}Y$, $W := U + \mathrm{i}V$. Assume that*

$$\begin{aligned} \mathrm{fl}(X+U) &= (X+U)(1+\varepsilon_1^+), \\ \mathrm{fl}(Y+V) &= (Y+V)(1+\varepsilon_2^+) \end{aligned}$$

where $\varepsilon_1^+, \varepsilon_2^+$ are independent of Z and W and with $\mathbb{E}(\varepsilon_j^+) = \mu_+ u$, $\mathbb{V}(\varepsilon_j^+) = \sigma_+^2 u^2$ for $j = 1, 2$. Define ε^+ by

$$\mathrm{fl}(Z+W) = (Z+W) + \varepsilon^+ \cdot |Z+W|.$$

Then we have:
(i) *The distribution of ε^+ is symmetric. In particular, $\mathbb{E}(\varepsilon^+) = 0$.*
(ii) *$\mathrm{Re}\,\varepsilon^+$, $\mathrm{Im}\,\varepsilon^+$ and $|Z+W|$ are uncorrelated.*
(iii) *We have $\mathbb{V}(\varepsilon^+) = (\mu_+^2 + \sigma_+^2)u^2$ and $\mathrm{Re}\,\varepsilon^+$ and $\mathrm{Im}\,\varepsilon^+$ have equal variances $\mathbb{V}(\mathrm{Re}\,\varepsilon^+) = \mathbb{V}(\mathrm{Im}\,\varepsilon^+)$.*

Proof. (i) From the definition we have

$$\varepsilon^+ = \frac{(X+U)\cdot\varepsilon_1^+ + \mathrm{i}(Y+V)\cdot\varepsilon_2^+}{\sqrt{(X+U)^2+(Y+V)^2}}.$$

By independence and the symmetry of the distributions of X, Y, U and V the distribution of $-\varepsilon^+ = \frac{(-X-U)\cdot\varepsilon_1^+ + \mathrm{i}(-Y-V)\cdot\varepsilon_2^+}{\sqrt{(-X-U)^2+(-Y-V)^2}}$ is the same as that of ε^+.
(ii) We have $\mathrm{Cov}\,(\mathrm{Re}\,\varepsilon^+, |Z+W|) = \mathbb{E}((X+U)\cdot\varepsilon_1^+) = \mathbb{E}(X+U)\cdot\mathbb{E}(\varepsilon_1^+) = 0\cdot\mu_+ u = 0$ and similarly for $\mathrm{Cov}\,(\mathrm{Im}\,\varepsilon^+, |Z+W|)$. Furthermore

$$\begin{aligned} &\mathrm{Cov}\,(\mathrm{Re}\,\varepsilon^+, \mathrm{Im}\,\varepsilon^+) \\ &\quad = \mathbb{E}\left(\frac{X+U}{\sqrt{(X+U)^2+(Y+V)^2}}\varepsilon_1^+ \cdot \frac{Y+V}{\sqrt{(X+U)^2+(Y+V)^2}}\varepsilon_2^+\right) \\ &\quad = \mu_+^2 u^2 \cdot \mathbb{E}\left(\frac{(X+U)(Y+V)}{(X+U)^2+(Y+V)^2}\right) \quad = \quad 0 \end{aligned}$$

again by symmetry.
(iii) Since $\mathbb{E}(\varepsilon^+) = 0$, we have

$$\begin{aligned} \mathbb{V}(\mathrm{Re}\,\varepsilon^+) &= \mathbb{E}\left(\frac{(X+U)^2}{(X+U)^2+(Y+V)^2}\cdot(\varepsilon_1^+)^2\right) \\ &= \mathbb{E}\left(\frac{(X+U)^2}{(X+U)^2+(Y+V)^2}\right)\cdot\mathbb{E}\big((\varepsilon_1^+)^2\big) = \frac{1}{2}\big(\mu_+^2+\sigma_+^2\big)u^2. \quad \blacksquare \end{aligned}$$

2.2 Remark. A more realistic model would be obtained by replacing the ε_j^+ from (1.1) with the fractional relative roundoff error η_j^+ from (1.2) in the assumptions in Proposition 2.1. Since by definition $\mathbb{E}((\eta_j^+)^2) = \mathbb{E}((\varepsilon_j^+)^2)\,\mathbb{E}((X+U)^2/\zeta(X+U)^2)$, we obtain

$$\mathbb{V}(\varepsilon^+) = \mathbb{E}\left(\frac{\zeta(X+U)^2+\zeta(Y+V)^2}{(X+U)^2+(Y+V)^2}\right)\mathbb{E}\left(\frac{(X+U)^2}{\zeta(X+U)^2}\right)(\mu_+^2+\sigma_+^2)u^2$$

instead of the result in (iii) above. The product of the two expected values (which depends on the distribution of the data) is very close to 1 for a large class of distributions. This justifies the use of (1.1) in the following.

In the next proposition we study complex multiplication. Here and in the following we assume that roundoff errors propagate without further alteration (i.e., they have no influence on further roundoff errors).

2.3 Proposition. *Let X, Y, U, V be real valued random variables such that X and Y have the same distribution which is symmetric around the origin, and set $Z := X + \mathrm{i}Y$, $W := U + \mathrm{i}V$. Assume that*

$$\mathrm{fl}(X \times U - Y \times V) = (X \times U - Y \times V)(1+\varepsilon_1^+) + (X \times U)\varepsilon_1^\times - (Y \times V)\varepsilon_2^\times$$

and

$$\mathrm{fl}(X \times V + Y \times U) = (X \times V + Y \times U)(1+\varepsilon_2^+) + (X \times V)\varepsilon_3^\times + (Y \times U)\varepsilon_4^\times$$

with $X, Y, W, \varepsilon_1^\times, \varepsilon_2^\times, \varepsilon_3^\times, \varepsilon_4^\times, \varepsilon_1^+, \varepsilon_2^+$ independent and $\mathbb{E}(\varepsilon_j^\times) = \mu_\times u$, $\mathbb{V}(\varepsilon_j^\times) = \sigma_\times^2 u^2$ $(j = 1,2,3,4)$ and $\mathbb{E}(\varepsilon_j^+) = \mu_+ u$, $\mathbb{V}(\varepsilon_j^+) = \sigma_+^2 u^2$ $(j = 1,2)$. If we define $\varepsilon^\times$ by

$$\mathrm{fl}(Z \times W) = Z \times W + \varepsilon^\times\,|Z \times W|,$$

then we have:
(i) *The distribution of $\varepsilon^\times$ is symmetric. In particular, $\mathbb{E}(\varepsilon^\times) = 0$.*
(ii) *$\mathrm{Re}\,\varepsilon^\times$, $\mathrm{Im}\,\varepsilon^\times$ and $|Z \times W|$ are uncorrelated.*
(iii) *$\mathrm{Re}\,\varepsilon^\times$ and $\mathrm{Im}\,\varepsilon^\times$ have the equal variances $\mathbb{V}(\mathrm{Re}\,\varepsilon^\times) = \mathbb{V}(\mathrm{Im}\,\varepsilon^\times)$ and $\mathbb{V}(\varepsilon^+) = (\sigma_\times^2 + \sigma_+^2 + \mu_\times^2 + \mu_+^2)u^2$.*

Proof. (i) This follows from

$$\mathrm{Re}\,\varepsilon^\times = \frac{XU}{\sqrt{X^2+Y^2}\sqrt{U^2+V^2}}\cdot(\varepsilon_1^\times+\varepsilon_1^+) - \frac{YV}{\sqrt{X^2+Y^2}\sqrt{U^2+V^2}}\cdot(\varepsilon_2^\times+\varepsilon_1^+)$$

and

$$\mathrm{Im}\,\varepsilon^\times = \frac{XV}{\sqrt{X^2+Y^2}\sqrt{U^2+V^2}}\cdot(\varepsilon_3^\times+\varepsilon_2^+) + \frac{YU}{\sqrt{X^2+Y^2}\sqrt{U^2+V^2}}\cdot(\varepsilon_4^\times+\varepsilon_2^+)$$

as well as the fact that the distribution of $\varepsilon^\times = \mathrm{Re}\,\varepsilon^\times + \mathrm{i}\,\mathrm{Im}\,\varepsilon^\times$ remains the same, if we replace (X, Y) by $(-X, -Y)$.

(ii) We have $\operatorname{Cov}(\operatorname{Re}\varepsilon^{\times}, |Z \times W|) = \mathbb{E}(XU\varepsilon_1^{\times} - YV\varepsilon_2^{\times} + (XU - YV)\varepsilon_1^{\times}) = \mathbb{E}(X)\mathbb{E}(U)\mathbb{E}(\varepsilon_1^{\times}) - \mathbb{E}(Y)\mathbb{E}(V)\mathbb{E}(\varepsilon_2^{\times}) + \big(\mathbb{E}(X)\mathbb{E}(U) - \mathbb{E}(Y)\mathbb{E}(V)\big)\mathbb{E}(\varepsilon_1^{+}) = 0$ and similarly for $\operatorname{Im}\varepsilon^{\times}$. Furthermore, $\operatorname{Cov}(\operatorname{Re}\varepsilon^{\times}, \operatorname{Im}\varepsilon^{\times})$ equals

$$\mathbb{E}\Big(\frac{\big((XU - YV)\varepsilon_1^{+} + XU\varepsilon_1^{\times} - YV\varepsilon_2^{\times}\big)\big((XV + YU)\varepsilon_2^{+} + XV\varepsilon_3^{\times} + YU\varepsilon_4^{\times}\big)}{(X^2 + Y^2)(U^2 + V^2)}\Big).$$

The terms going with $\mathbb{E}(\varepsilon_1^{+})\mathbb{E}(\varepsilon_2^{+})$ are

$$\begin{aligned}&\mathbb{E}\Big(\frac{X^2}{X^2+Y^2}\Big)\mathbb{E}\Big(\frac{UV}{U^2+V^2}\Big) \;+\; \mathbb{E}\Big(\frac{XY}{X^2+Y^2}\Big)\mathbb{E}\Big(\frac{U^2}{U^2+V^2}\Big)\\ &-\; \mathbb{E}\Big(\frac{XY}{X^2+Y^2}\Big)\mathbb{E}\Big(\frac{V^2}{U^2+V^2}\Big) \;-\; \mathbb{E}\Big(\frac{Y^2}{X^2+Y^2}\Big)\mathbb{E}\Big(\frac{UV}{U^2+V^2}\Big)\end{aligned}$$

and by our assumptions the first and the last terms cancel, whereas the two middle terms vanish because of $\mathbb{E}\big(\frac{XY}{X^2+Y^2}\big) = 0$. Here the symmetry of the distribution of X is used. The same terms also occur in conjunction with $\mathbb{E}(\varepsilon_j^{\times})\mathbb{E}(\varepsilon_k^{\times}) = \mu_{\times}^2 u^2$ (where $j = 1, 2$, $k = 3, 4$) and with the mixed terms $\mathbb{E}(\varepsilon_j^{+})\mathbb{E}(\varepsilon_k^{\times}) = \mu_{+}\mu_{\times} u^2$ (where $j = 1, k = 3, 4$ as well as $j = 2, k = 1, 2$) and therefore $\operatorname{Cov}(\operatorname{Re}\varepsilon^{\times}, \operatorname{Im}\varepsilon^{\times}) = 0$.
(iii) This is a similar calculation as in (ii). ■

2.4 Remark. It is possible to use only three multiplications (but five additions) by defining $\alpha := (X+Y)\times(U-V)$, $\beta := X \times V$, $\gamma := Y \times U$ and using the identity

$$(X + \mathrm{i}Y) \times (U + \mathrm{i}V) = (\alpha + (\beta - \gamma)) + \mathrm{i}(\beta + \gamma).$$

Whereas the usual complex multiplication is based on the formula

$$\begin{pmatrix} XU - YV \\ XV + YU \end{pmatrix} = \begin{pmatrix} X & -Y \\ Y & X \end{pmatrix} \begin{pmatrix} U \\ V \end{pmatrix}$$

with an almost orthogonal matrix, the three–plus–five algorithm relies on

$$\begin{pmatrix} XU - YV \\ XV + YU \end{pmatrix} = \begin{pmatrix} 1 & 1 & -1 \\ 0 & 1 & 1 \end{pmatrix} \begin{pmatrix} X+Y & 0 & 0 \\ 0 & X & 0 \\ 0 & 0 & Y \end{pmatrix} \begin{pmatrix} 1 & -1 \\ 0 & 1 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} U \\ V \end{pmatrix}$$

with non–orthogonal matrices. Therefore it is not surprising, that in terms of numerical stability this method is worse than the usual method. In fact, under the same assumptions on the data as in Proposition 2.3 one can prove that the relative error

$$\begin{aligned}\varepsilon^{\times\prime} &:= \frac{\mathrm{fl}((\alpha + (\beta - \gamma)) + \mathrm{i}(\beta + \gamma)) - (X + \mathrm{i}Y)(U + \mathrm{i}V)}{|(X + \mathrm{i}Y)(U + \mathrm{i}V)|}\\ &= \frac{\alpha(\varepsilon_1^{+} + \varepsilon_2^{+} + \varepsilon_1^{\times} + \varepsilon_4^{+}) + \beta\varepsilon_2^{\times} - \gamma\varepsilon_3^{\times} + (\beta - \gamma)(\varepsilon_3^{+} + \varepsilon_4^{+})}{|(X + \mathrm{i}Y)(U + \mathrm{i}V)|}\\ &\quad + \mathrm{i}\,\frac{\beta\varepsilon_2^{\times} + \gamma\varepsilon_3^{\times} + (\beta + \gamma)\varepsilon_5^{+}}{|(X + \mathrm{i}Y)(U + \mathrm{i}V)|} \;+\; \mathcal{O}(u^2)\end{aligned}$$

(where $\varepsilon_1^\times, \ldots, \varepsilon_5^+$ are the relative roundoff errors from the successive computations) satisfies $\mathbb{E}(\varepsilon^{\times\prime}) = 0$,

$$\begin{aligned}
\mathbb{V}(\operatorname{Re}\varepsilon^{\times\prime}) &= \Big(3\sigma_+^2 + \frac{3}{2}\sigma_\times^2 \;+\; 5\mu_+^2 + 3\mu_+\mu_\times + \frac{1}{2}\mu_\times^2\Big)u^2 \;+\; \mathcal{O}(u^3),\\
\mathbb{V}(\operatorname{Im}\varepsilon^{\times\prime}) &= \frac{1}{2}\big(\sigma_+^2 + \sigma_\times^2 \;+\; \mu_+^2 + \mu_\times^2\big)u^2 \;+\; \mathcal{O}(u^3),\\
\operatorname{Cov}(\operatorname{Re}\varepsilon^{\times\prime}, \operatorname{Im}\varepsilon^{\times\prime}) &= \mathcal{O}(u^3),
\end{aligned}$$

and so in particular $\mathbb{E}(|\varepsilon^{\times\prime}|^2) = \big(\frac{7}{2}\sigma_+^2 + 2\sigma_\times^2 \;+\; \frac{11}{2}\mu_+^2 + 3\mu_+\mu_\times + \mu_\times^2\big)u^2 + \mathcal{O}(u^3)$ is larger than for the usual computation of the product of two complex numbers (see Proposition 2.3 (iii)).

By reversing the role of real and imaginary parts we can obtain an algorithm with the same complexity and similar error variance. Yet another method with the same complexity is described in [10], p. 73 and based on $XU - YV = (X + Y)U - Y(U + V)$, $XV + YU = (X + Y)U + X(V - U)$. It leads to a complex roundoff error with equal variances of real and imaginary parts but with covariance $\frac{1}{2}\big(\sigma_+^2 + \sigma_\times^2 + \mu_+^2 + \mu_\times^2\big)u^2 + \mathcal{O}(u^3)$. Therefore they are *not* uncorrelated as in the case of all the other algorithms. The total variance of the relative error is $(3\sigma_+^2 + 2\sigma_\times^2 \;+\; 3\mu_+^2 + 2\mu_\times^2)u^2 + \mathcal{O}(u^3)$.

2.5 Remark. In the following diagrams of Figure 3 we show the results of 10000 simulated multiplications of complex numbers using the methods from Proposition 2.3 (left) and Remark 2.4 (right diagram).

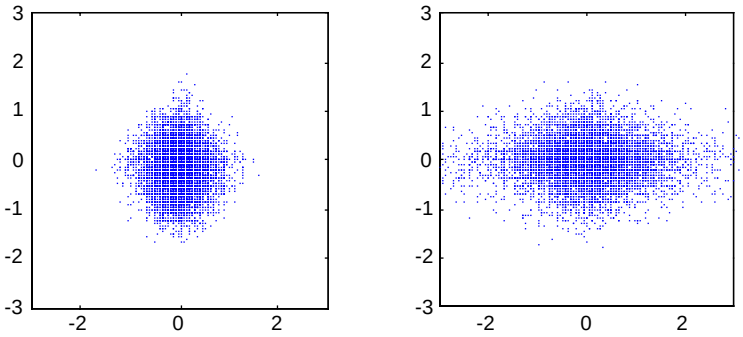


Figure 3: Distribution of roundoff errors $\varepsilon^\times/u$ for different methods of complex multiplication

As can be seen from these diagrams the distribution of the complex valued relative roundoff error $\varepsilon^\times$ is not exactly rotation symmetric, even when Z and W are. This is also true for ε^+.

If one of the factors of the product is a real number (or purely imaginary) then only two real multiplications are needed to calculate the product. As a consequence, the roundoff errors are smaller than in the general situation.

2.6 Proposition. *Let X, Y, U be independent real valued random variables such that the distributions of X and Y are equal and symmetric around the origin, and set $Z := X + \mathrm{i}Y$. Assume that*

$$\begin{aligned}
\operatorname{fl}(U \times X) &= (U \times X)(1 + \varepsilon_1^\times)\\
\operatorname{fl}(U \times Y) &= (U \times Y)(1 + \varepsilon_2^\times)
\end{aligned}$$

with $\varepsilon_1^\times, \varepsilon_2^\times$ *independent of* Z *and* U *and with* $\mathbb{E}(\varepsilon_j^\times) = \mu_\times u$, $\mathbb{V}(\varepsilon_j^\times) = \sigma_\times^2 u^2$ *(for* $j = 1, 2$*), and define* $\varepsilon^\times$ *by*

$$\mathrm{fl}(U \times Z) = (U \times Z) + \varepsilon^\times |U \times Z|.$$

Then we have:
(i) *The distribution of* $\varepsilon^\times$ *is symmetric. In particular* $\mathbb{E}(\varepsilon^\times) = 0$.
(ii) $\operatorname{Re}\varepsilon^\times$, $\operatorname{Im}\varepsilon^\times$ *and* $|U \times Z|$ *are uncorrelated.*
(iii) $\operatorname{Re}\varepsilon^\times$ *and* $\operatorname{Im}\varepsilon^\times$ *have equal variances* $\mathbb{V}(\operatorname{Re}\varepsilon^\times) = \mathbb{V}(\operatorname{Im}\varepsilon^\times)$, *and* $\mathbb{V}(\varepsilon^\times) = (\mu_\times^2 + \sigma_\times^2)u^2$.

The proof is similar to the preceding ones and will be omitted.

Finally we study division in $\mathbb{C}$. We will restrict our attention to the roundoff errors in the computation of the inverse $1/Z$, where Z is a complex variable such that $Z \neq 0$ almost surely. Unlike the cases of addition and multiplication the variance of the roundoff error is only approximately independent of the distribution of the data.

2.7 Proposition. *Let* X, Y *be independent with equal and symmetric distributions such that* $P\{X = 0\} = 0$, *and set* $Z := X + \mathrm{i}Y$. *Assume that*

$$\mathrm{fl}\Big(\frac{X}{X^2+Y^2}\Big) = \frac{X}{X^2\varepsilon_1^\times + Y^2\varepsilon_2^\times + (X^2+Y^2)(1+\varepsilon_1^+)}(1+\varepsilon_1^\div),$$

$$\mathrm{fl}\Big(\frac{Y}{X^2+Y^2}\Big) = \frac{Y}{X^2\varepsilon_1^\times + Y^2\varepsilon_2^\times + (X^2+Y^2)(1+\varepsilon_1^+)}(1+\varepsilon_2^\div)$$

with $X, Y, \varepsilon_1^\times, \varepsilon_2^\times, \varepsilon_1^+, \varepsilon_1^\div, \varepsilon_2^\div$ *independent,* $\mathbb{E}(\varepsilon_j^\times) = \mu_\times u$, $\mathbb{V}(\varepsilon_j^\times) = \sigma_\times^2 u^2$, $\mathbb{E}(\varepsilon_j^+) = \mu_+ u$, $\mathbb{V}(\varepsilon_j^+) = \sigma_+^2 u^2$ *for* $j = 1, 2$ *and* $\mathbb{E}(\varepsilon_1^\div) = \mu_\div u$, $\mathbb{V}(\varepsilon_1^\div) = \sigma_\div^2 u^2$. *If the relative complex roundoff error* $\varepsilon^\div$ *is defined by*

$$\mathrm{fl}\Big(\frac{1}{Z}\Big) \quad = \quad \frac{1}{Z} \; + \; \frac{\varepsilon^\div}{|Z|}$$

then we have:
(i) *The distribution of* $\varepsilon^\div$ *is symmetric. In particular,* $\mathbb{E}(\varepsilon^\div) = 0$.
(ii) $\operatorname{Cov}(\operatorname{Re}\varepsilon^\div, \operatorname{Im}\varepsilon^\div)$, $\operatorname{Cov}(\operatorname{Re}\varepsilon^\div, 1/|Z|)$, $\operatorname{Cov}(\operatorname{Im}\varepsilon^\div, 1/|Z|) = \mathcal{O}(u^3)$.
(iii) *Let* $\kappa := \mathbb{E}\big((X^4+Y^4)/(X^2+Y^2)^2\big)$. *Then* $\mathbb{V}(\operatorname{Re}\varepsilon^\div) = \mathbb{V}(\operatorname{Im}\varepsilon^\div)$ *and*

$$\mathbb{V}(\varepsilon^\div) = (\kappa\sigma_\times^2 + \sigma_\div^2 + \sigma_+^2 \; + \; \mu_\times^2 + \mu_\div^2 + \mu_+^2 + 2\mu_\times\mu_+ - 2(\mu_\times + \mu_+)\mu_\div)u^2 + \mathcal{O}(u^3).$$

Proof. Let $\delta := \frac{X^2\varepsilon_1^\times + Y^2\varepsilon_2^\times}{X^2+Y^2} + \varepsilon_1^+$. Then it follows from the definition of $\varepsilon^\div$ that $\operatorname{Re}\varepsilon^\div = \frac{X}{\sqrt{X^2+Y^2}}\big(\varepsilon_1^\div - \delta\big) + \mathcal{O}(u^2)$ and $\operatorname{Im}\varepsilon^\div = \frac{-Y}{\sqrt{X^2+Y^2}}\big(\varepsilon_2^\div - \delta\big) + \mathcal{O}(u^2)$.
(i) This is a consequence of the symmetry of the distributions of X and Y.
(ii) The covariances satisfy $\operatorname{Cov}(\operatorname{Re}\varepsilon^\div, \operatorname{Im}\varepsilon^\div) = \mathbb{E}\big(\frac{-XY}{X^2+Y^2}(\varepsilon_1^\div - \delta)(\varepsilon_2^\div - \delta)\big) + \mathcal{O}(u^3)$, $\operatorname{Cov}(\operatorname{Re}\varepsilon^\div, 1/|Z|) = \mathbb{E}\big(\frac{X}{X^2+Y^2}(\varepsilon_1^\div - \delta)\big) + \mathcal{O}(u^3)$, and the expected values are 0 by the symmetry of the distributions of X and Y.

(iii) Obviously, $\mathbb{V}(\operatorname{Re}\varepsilon^{\div}) = \mathbb{E}\big(\frac{X^2}{X^2+Y^2}(\varepsilon_1^{\div}-\delta)^2\big) + \mathcal{O}(u^3) = \mathbb{V}(\operatorname{Im}\varepsilon^{\div})$. Since $\mathbb{E}(X^2(X^4+Y^4)/(X^2+Y^2)^3) = \frac{\kappa}{2}$, the result now follows from

$$\begin{aligned}
\mathbb{E}\bigg(\frac{X^2}{X^2+Y^2}(\varepsilon_1^{\div})^2\bigg) &= \mathbb{E}\bigg(\frac{X^2}{X^2+Y^2}\bigg)\mathbb{E}\big((\varepsilon_1^{\div})^2\big) = \frac{1}{2}(\sigma_{\div}^2+\mu_{\div}^2)u^2,\\
\mathbb{E}\bigg(\frac{X^2}{X^2+Y^2}\,\varepsilon_1^{\div}\delta\bigg) &= \mathbb{E}\bigg(\frac{X^2}{X^2+Y^2}\Big(\frac{X^2\varepsilon_1^{\times}+Y^2\varepsilon_2^{\times}}{X^2+Y^2}+\varepsilon_1^{+}\Big)\bigg)\mu_{\div}u\\
&= \frac{1}{2}(\mu_{\times}+\mu_{+})\mu_{\div}u^2,\\
\mathbb{E}\bigg(\frac{X^2}{X^2+Y^2}\,\delta^2\bigg) &= \Big(\frac{\kappa}{2}\sigma_{\times}^2+\frac{1}{2}\mu_{\times}^2+\mu_{\times}\mu_{+}+\frac{1}{2}(\sigma_{+}^2+\mu_{+}^2)\Big)u^2. \qquad \blacksquare
\end{aligned}$$

The constant κ depends on the distribution of X and Y and takes values in the interval $\big[\frac{1}{2},1\big]$. For the standard normal distribution we obtain $\kappa = \frac{3}{4}$.

2.8 Conclusion. It follows from Propositions 2.1, 2.3, 2.6 and 2.7 that the roundoff error $\Delta^{\circ} := \operatorname{fl}(X\circ Y) - X\circ Y = |X\circ Y|\,\varepsilon^{\circ}$ satisfies

$$\mathbb{E}(\Delta^{\circ}) = 0, \qquad \mathbb{V}(\Delta^{\circ}) = \mathbb{E}(|X\circ Y|^2)\,\mathbb{E}\big(|\varepsilon^{\circ}|^2\big) \tag{2.1}$$

for all arithmetic operations $\circ \in \{+,-,\times,\div\}$.

This is essentially a consequence of the assumption that the data has expectation zero and is independent of the relative error. It holds even when $\mu^{\circ} \neq 0$. But (2.1) is valid even without the assumption on the expectation of the data, if we know $\mathbb{E}(\varepsilon^{\circ}) = \mu^{\circ} = 0$ instead.

In simulations, the standard deviations of the roundoff errors ε^{+} and $\varepsilon^{\times}$ are up to 10% larger than to be expected from Propositions 2.1, 2.3, 2.6 and 2.7. Partially this is due to the effect observed in Remark 2.2, partially to the fact that, although the result of an operation and the roundoff error are uncorrelated, our stronger assumption of independence is only approximately satisfied.

3. Stochastic roundoff model

Motivated by the propositions in the Section 2 and Remarks 1.9 and 2.8 we will describe the roundoff errors in operations with real or complex numbers by the following model:

3.1 Model. For each roundoff error $\Delta^{\circ} = \operatorname{fl}(X\circ Y) - X\circ Y = |X\circ Y|\cdot\varepsilon^{\circ}$ occuring during the computation of an elementary arithmetic operation $\circ$ in $\mathbb{K} = \mathbb{R}$ or $\mathbb{C}$ we make the following assumptions:

(M1) The relative error ε° is independent of the data (and in particular independent of the operands).

(M2) The relative error has expectation $\mathbb{E}(\varepsilon^{\circ}) = 0$ and variance $\mathbb{V}(\varepsilon^{\circ}) = \sigma_{\mathbb{K},\circ}^2 u^2$. Therefore the total error Δ° has expectation and variance

$$\mathbb{E}(\Delta^{\circ}) = 0, \qquad \mathbb{V}(\Delta^{\circ}) = \|\!|X\circ Y|\!\|^2\,\sigma_{\mathbb{K},\circ}^2 u^2\,, \tag{3.1}$$

where $\|\!|Z|\!\| := \mathbb{E}(|Z|^2)^{1/2}$ denotes the L^2-norm of a random variable Z. The constant $\sigma_{\mathbb{K},\circ}$ depends on the operation $\circ$ and the number field $\mathbb{K}$. It follows from the results

in Sections 1 and 2 that $\sigma^2_{\mathbb{R},+} = \sigma^2_{\mathbb{C},+} = \sigma^2_+ + \mu^2_+$, $\sigma^2_{\mathbb{R},\times} = \sigma^2_\times + \mu^2_\times$ and $\sigma^2_{\mathbb{C},\times} = \sigma^2_\times + \sigma^2_+ + \mu^2_\times + \mu^2_+$, where the basic constants $\sigma_\circ$ and $\mu_\circ$ ($\circ = +$ or $\times$) are defined in (1.3) and (1.4). In the case $\mathbb{K} = \mathbb{C}$ of complex operands we further assume that $\operatorname{Re}\varepsilon^\circ$ and $\operatorname{Im}\varepsilon^\circ$ are uncorrelated and $\mathbb{V}(\operatorname{Re}\varepsilon^\circ) = \mathbb{V}(\operatorname{Im}\varepsilon^\circ)$. This is motivated by parts (ii) and (iii) of Propositions 2.1, 2.3, 2.6 and 2.7.

(M3) We assume that all relative roundoff errors ε_j occuring within one step of an algorithm are independent. This implies that all the *total* roundoff errors Δ_j are uncorrelated. In fact, if the vector $\boldsymbol{X}$ denotes the data, we can write each roundoff error in the form $\Delta_j = \varepsilon_j f_j(\boldsymbol{X})$ and therefore $\operatorname{Cov}(\Delta_j, \Delta_k) = \mathbb{E}(\varepsilon_j f_j(\boldsymbol{X})\overline{\varepsilon_k}\,\overline{f_k(\boldsymbol{X})}) = \mathbb{E}(\varepsilon_j)\mathbb{E}(\overline{\varepsilon_k})\mathbb{E}(f_j(\boldsymbol{X})\overline{f_k(\boldsymbol{X})}) = 0$.

(M4) We furthermore assume that all the input variables are uncorrelated with known expectations and variances. We then also know the $\mathcal{L}^2$-norm

$$\|\!|X|\!\| = \big(\mathbb{V}(X) + |\mathbb{E}(X)|^2\big)^{1/2}.$$

If the data is complex valued, the real and imaginary parts have to be uncorrelated with equal variances.

(M5) Finally we assume that roundoff errors occuring in one step of the algorithm are propagated linearly through the later steps. In particular, we neglect terms of order $\mathcal{O}(u^2)$.

3.2 Remarks. The last part of condition (M2) in the complex case is obviously equivalent to $\mathbb{E}([\varepsilon^\circ]^2) = \mathbb{E}([\operatorname{Re}\varepsilon^\circ]^2 - [\operatorname{Im}\varepsilon^\circ]^2) + 2\mathrm{i}\,\mathbb{E}(\operatorname{Re}\varepsilon^\circ \operatorname{Im}\varepsilon^\circ) = 0$. Similarly, the last part of (M4) is equivalent to $\mathbb{E}(X^2) = [\mathbb{E}(X)]^2$ for each component of the data. By a simple calculation we see that this property is preserved for every linear combination of uncorrelated components. Therefore, by (M4), for any algorithm proceeding in a series of matrix vector products, the condition on the real and imaginary part holds for all intermediary results.

If an algorithm consists of several steps, then the assumption that the data are uncorrelated has to be checked for the intermediary result used in each step. For many algorithms each step consists in the multiplication of the data with a unitary matrix (or a multiple thereof). If the data are uncorrelated and all variances are equal, then these two properties are preserved in each step. Note that our model cannot be applied directly and has to be adapted if one of the algorithms in Remark 2.4 for the complex multiplication is used. Nonunitary algorithms often show more irregular roundoff error distributions (see for example [2], Figure 3 and Remarks 2.5 and 3.10 below).

In condition (M2) of Model 3.1 we assume that the constants $\sigma_{\mathbb{K},\circ}$ do not depend on the distribution of the data. For multiplication (and to a large extent for division) this is motivated by Propositions 1.4 and 2.3, whereas for summation we argue that the distribution of the exponent and sign of data (which essentially determines the roundoff errors) is well approximated by what would be obtained for standard Gaussian data – at least for most of the intermediary results of an algorithm.

Instead of the independence assumptions (M1) in our Model 3.1 it is also possible to use the simpler but less realistic condition that the absolute error Δ° is uncorrelated with the data. This makes the calculations much easier and in the end leads to the same results. As the result of an average case study of a particular algorithm we usually state the variances of the components of the roundoff error vector and

its covariance structure. For specific values of the data it is quite possible that the roundoff error is much larger or smaller.

In many cases, in addition to (M4) we will impose the assumption that the data has expectation zero. Then the norm $\|\!|X|\!\|$ is equal to the variance $\mathbb{V}(Z)$. The general formula

$$\|\!|X+Y|\!\|^2 = \|\!|X|\!\|^2 + \|\!|Y|\!\|^2 + 2\,\mathrm{Re}\big(\mathbb{E}(X)\mathbb{E}(\overline{Y})\big) \tag{3.2}$$

for uncorrelated variables of arbitrary expectation then takes the simpler form $\|\!|X+Y|\!\|^2 = \|\!|X|\!\|^2 + \|\!|Y|\!\|^2$, which alleviates the computation of the norm of the data in subsequent steps of an algorithm. The assumption that the data has expectation 0 is valid for many algorithms (e.g. JPEG). However, the assumption that the data are uncorrelated is essential for our results, and so the analysis of a compression algorithm, where it is the main point that the data is *not* uncorrelated, would require a more complicated model.

3.3 Remark. It is interesting to note that for independent U, X, Y in $\mathbb{K} = \mathbb{R}$ or $\mathbb{C}$ with $\mathbb{E}(X) = 0$ or $\mathbb{E}(Y) = 0$, the variance of the roundoff error in the calculation of $UX + UY = U(X+Y)$ is $\sigma^2_{\mathbb{K},\times}(\|\!|UX|\!\|^2 + \|\!|UY|\!\|^2)u^2 + \sigma^2_{\mathbb{K},+}\|\!|UX+UY|\!\|^2u^2$ if we calculate the left hand side (using two multiplications and one addition), and the variance is $\sigma^2_{\mathbb{K},\times}\|\!|U(X+Y)|\!\|^2u^2 + \|\!|U|\!\|^2\sigma^2_{\mathbb{K},+}\|\!|X+Y|\!\|^2u^2$ for the roundoff error of the right hand side. Although the number of multiplications is different, the two variances are equal. The same holds for the use of the distributive law with more than two summands. Therefore with respect to roundoff errors in a matrix–vector multiplication, it matters little if we exploit the equality between matrix entries of the same row to reduce the number of multiplications (as long as the data have expectation 0).

As an auxiliary result needed for the study of the roundoff error of a matrix–vector calculation, we investigate properties of the "inner product" $\boldsymbol{x}^{\mathrm{T}}\boldsymbol{y} = \sum_{j=0}^{n-1} x_j y_j$ of two vectors $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{C}^n$.

3.4 Lemma. *Let $\boldsymbol{X} = (X_j)_{j=0}^{n-1}$ and $\boldsymbol{Y} = (Y_j)_{j=0}^{n-1}$ be independent random vectors in $\mathbb{K} = \mathbb{R}$ or $\mathbb{C}$ with uncorrelated components satisfying condition (M4) and $\mathbb{E}(Y_j) = 0$ for all $j = 0, \dots, n-1$. Then we have*

(i) *The roundoff error Δ occuring in the recursive calculation of*

$$\sum_{j=0}^{n-1} X_jY_j = (((X_0Y_0 + X_1Y_1) + X_2Y_2) + \dots) + X_{n-1}Y_{n-1}$$

has expectation zero and variance

$$\begin{aligned}\mathbb{V}(\Delta) = &\Big((n-1)\|\!|X_0|\!\|^2\|\!|Y_0|\!\|^2 + \sum_{j=1}^{n-1}(n-j)\|\!|X_j|\!\|^2\|\!|Y_j|\!\|^2\Big)\,\sigma^2_{\mathbb{K},+}u^2 \\ &+ \sum_{j=0}^{n-1}\|\!|X_j|\!\|^2\|\!|Y_j|\!\|^2\,\sigma^2_{\mathbb{K},\times}u^2\,.\end{aligned} \tag{3.3}$$

(ii) *For cascade summation the variance of the roundoff error is*

$$\mathbb{V}(\Delta) = \Big(\lceil\log_2 n\rceil\sum_{j=0}^{n-1}\|\!|X_j|\!\|^2\|\!|Y_j|\!\|^2 - r\Big)\sigma^2_{\mathbb{K},+}u^2 + \sum_{j=0}^{n-1}\|\!|X_j|\!\|^2\|\!|Y_j|\!\|^2\sigma^2_{\mathbb{K},\times}u^2$$

where

$$r := \sum_{\substack{s=0 \\ \lceil n/2^s \rceil \text{ odd}}}^{\lceil \log_2 n \rceil - 2} r_s$$

is the sum of the norms r_s of the term which is left out in stage s of the cascade.

Proof. (i) Since $\mathbb{E}(X_j Y_j) = 0$, this follows as in [15], Lemma 5.1.
(ii) The calculation of the products $X_j Y_j$ gives rise to the roundoff errors $\Delta_j^\times = \|X_j\|\|Y_j\|\varepsilon_j^\times$ of expectation 0 and variance $\|X_j\|^2\|Y_j\|^2\,\sigma_{\mathbb{K},\times}^2 u^2$ by (3.1). By (M3) the variance of the sum of roundoff errors is equal to the sum of variances. At stage $s = 0, \dots, \lceil \log_2 n \rceil - 1$ of the cascade the $\lceil n/2^s \rceil$ summands are grouped in pairs. If $\lceil n/2^s \rceil$ is odd, one term, which has variance r_s, is carried to the next stage unchanged. The remaining pairs are added; the computation of the sum $\sum_{j\in S} X_j Y_j$ with $S \subseteq \{0, \dots, n-1\}$ produces a roundoff error of variance $\sigma_{\mathbb{K},+}^2 \|\sum_{j\in S} X_j Y_j\|_2^2\, u^2 = \sigma_{\mathbb{K},+}^2 \sum_{j\in S} \|X_j\|_2^2 \|Y_j\|_2^2\, u^2$. Hence by (3.2) the variance of sum of the roundoff errors at this stage equals $\big(\sum_{j=0}^{n-1} \|X_j\|^2\|Y_j\|^2 - r_s\big)\sigma_{\mathbb{K},+}^2 u^2$. ∎

3.5 Remark. As in the worst case study (see [7], p. 90) the size of the roundoff error depends strongly on the order of the summation, and it is clear from (i) that more accurate results are obtained, if we start the summation with terms of small norms. Likewise, the number r in (ii) depends on the order in which the summands in the cascade are combined. In order to obtain a small variance of the roundoff error, we try to make r large. For example, if all $\|X_j\|\|Y_j\|$ have the same value ρ, then at all stages s of the cascade where terms are left out, i.e., where $\lceil n/2^s \rceil$ is odd, we can make r_s as large as $2^s\rho^2$, and therefore the maximum is easily calculated as $r = \big(2^{\lceil \log_2 n \rceil} - n\big)\rho^2$. This leads to a variance of the roundoff error $\mathbb{V}(\Delta) = \big(n\lceil \log_2 n \rceil + n - 2^{\lceil \log_2 n \rceil}\big)\rho^2\sigma_{\mathbb{K},+}^2 u^2 + n\rho^2\sigma_{\mathbb{K},\times}^2 u^2$.

If the expected values of the data are not 0, the situation is more complicated. We will analyse the simplest case only.

3.6 Lemma. *Let $n = 2^t$ and $\boldsymbol{X} = (X_j)_{j=0}^{n-1}$ and $\boldsymbol{Y} = (Y_j)_{j=0}^{n-1}$ be independent random vectors with uncorrelated components. Then the roundoff error Δ in the calculation of $\sum_{j=0}^{n-1} X_j Y_j$ by cascade has variance*

$$\begin{aligned}\mathbb{V}(\Delta) = &\sum_{j=0}^{n-1} \|X_j\|^2\|Y_j\|^2 (t\sigma_{\mathbb{K},+}^2 + \sigma_{\mathbb{K},\times}^2)u^2 \\ &+ 2\sum_{j<k} c(j,k)\,\mathrm{Re}\big(\mathbb{E}(X_j)\mathbb{E}(Y_j)\overline{\mathbb{E}(X_k)\mathbb{E}(Y_k)}\big)\sigma_{\mathbb{K},+}^2 u^2 \qquad (3.4)\end{aligned}$$

where $c(j,k) := t + 1 - \min\{s = 1, \dots, t : \lfloor \frac{j}{2^s} \rfloor = \lfloor \frac{k}{2^s} \rfloor\}$.

Proof. Now we use (3.2) to evaluate

$$\|\sum_{j\in S} X_j Y_j\|_2^2 = \sum_{j\in S} \|X_j\|_2^2\|Y_j\|_2^2 + 2\sum_{j<k\in S} \mathrm{Re}\big(\mathbb{E}(X_j Y_j)\overline{\mathbb{E}(X_k Y_k)}\big)$$

for $S \subseteq \{0, \dots, n-1\}$. Since every compound term $\mathbb{E}(X_j Y_j)\mathbb{E}(\overline{X_k Y_k})$ arises exactly $c(j,k)$ times in the cascade, the result follows. ∎

3.7 Example. If all products $\mathbb{E}(X_j)\mathbb{E}(Y_j)$ are equal to μ, then the double sum in (3.4) is equal to $n(2n-2-t)|\mu|^2\sigma_{\mathbb{K},+}^2u^2$. Here we have used the identity $\sum_{j<k} c(j,k) = \frac{1}{2}\sum_{s=1}^{t} 2^s(2^s-1)\frac{n}{2^s} = \frac{n}{2}(2n-2-t)$.

In the following result we will study the roundoff errors from a matrix–vector product $\boldsymbol{AX}$. In addition to the arithmetical roundoff errors studied above, we also have to consider precomputation errors in the coefficients $A_{j,k}$ of the matrix $\boldsymbol{A}$. Here we have to take into account that when several entries of the matrix are equal, then the corresponding precomputation errors are also equal, as well as the roundoff errors from identical multiplications. More generally, we study the consequences of

$$\mathrm{fl}((\alpha x)y) - (\alpha x)y = \alpha\big(\mathrm{fl}(x\,y) - x\,y\big) \quad \text{for } \alpha \in \{0,1,-1,\mathrm{i},-\mathrm{i}\} \text{ and } x,y \in \mathbb{C}. \tag{3.5}$$

3.8 Theorem. *Let* $\boldsymbol{A} = (A_{j,k})_{j,k=0}^{n-1} = \boldsymbol{A}^{(0)} + \sum_{l=1}^{m} a_l\boldsymbol{A}^{(l)}$ *with* $a_l \in \mathbb{C}$ *for* $l \geq 1$, $A_{j,k}^{(l)} \in \{0,1,-1,\mathrm{i},-\mathrm{i}\}$ *and* $A_{j,k}^{(l)}A_{j,k}^{(l')} = 0$ *for all* $0 \leq l \neq l' \leq m$, $j,k = 0,\ldots,n-1$. *Let*

$$\hat{\boldsymbol{A}} = \boldsymbol{A}^{(0)} + \sum_{l=1}^{m}(a_l + \delta_l^A)\boldsymbol{A}^{(l)}$$

be the matrix with the precomputed entries of $\boldsymbol{A}$ *and let* $\boldsymbol{X}$ *have uncorrelated components* X_l *of variance* ρ^2 *and expectation 0. We assume that the products* $A_{j,k}^{(0)}X_k$ *are computed exactly, that the conditions in Model* 3.1 *are satisfied and that intermediary results of summations are only used once.*

(i) *Then* $\boldsymbol{\Delta} := \mathrm{fl}(\hat{\boldsymbol{A}}\boldsymbol{X}) - \boldsymbol{AX}$ *has the covariance matrix*

$$\mathrm{Cov}\,(\boldsymbol{\Delta}) = \Sigma^+u^2 + \rho^2\sum_{l=1}^{m}\big(\mathbb{E}\big(|\delta_l^A|^2\big) + |a_l|^2\sigma_{\mathbb{K},\times}^2u^2\big)\boldsymbol{A}^{(l)}\boldsymbol{A}^{(l)\,*} \tag{3.6}$$

where $\boldsymbol{\Sigma}^+$ *is the diagonal matrix with entries*

$$\Sigma_{j,j}^+ = \big((n-1)|A_{j,0}|^2 + \sum_{k=1}^{n-1}(n-l)|A_{j,k}|^2\big)\,\sigma_{\mathbb{K},+}^2\rho^2$$

in the case of recursive summation. Each $\sigma_{\mathbb{K},\times}^2$ *is either* $\sigma_{\mathbb{R},\times}^2$ *or* $\sigma_{\mathbb{C},\times}^2$ *depending on whether* $a_l \in \mathbb{R}\cup\mathrm{i}\mathbb{R}$ *or* $a_l \in \mathbb{C}\setminus(\mathbb{R}\cup\mathrm{i}\mathbb{R})$. *For cascade summation,* $\Sigma_{j,j}^+ = \big(\lceil\log_2 n\rceil\sum_{k=0}^{n-1}|A_{j,k}|^2 - r(j)\big)\sigma_{\mathbb{K},+}^2\rho^2$ *with* $r(j) \geq 0$ *defined as the number* r *in Lemma* 3.4 *(ii) with* $\boldsymbol{Y}$ *equal to row* j *of* $\boldsymbol{A}$.

(ii) *The norm of the error satisfies* $\mathbb{E}(\|\boldsymbol{\Delta}\|_2^2) = \bar{k}_n^2\mathbb{E}(\|\boldsymbol{X}\|_2^2)u^2$ *with the constant*

$$\bar{k}_n^2 := \frac{1}{n}\big(\lceil\log_2 n\rceil\|\boldsymbol{A}\|_F^2 - r\big)\sigma_{\mathbb{K},+}^2 + \frac{1}{n}\sum_{l=1}^{m}\big(\mathbb{E}\big(|\delta_l^A|^2/u^2\big) + |a_l|^2\sigma_{\mathbb{K},\times}^2\big)\|\boldsymbol{A}^{(l)}\|_F^2 \tag{3.7}$$

in case of cascade summation. Here $\|\boldsymbol{A}\|_F$ *denotes the Frobenius norm of the matrix* $\boldsymbol{A}$ *and* $r := \sum_{j=0}^{n-1} r(j)$. *If* n *is a power of* 2, *then* $r = 0$. *When all constants* a_l *are in* $\mathbb{R}\cup\mathrm{i}\mathbb{R}$ *or all are not in* $\mathbb{R}\cup\mathrm{i}\mathbb{R}$, *then all* $\sigma_{\mathbb{K},\times}^2$ *in* (3.6) *are equal and formula* (3.7) *simplifies to*

$$\bar{k}_n^2 = \frac{1}{n}\Big(\big(\lceil\log_2 n\rceil\|\boldsymbol{A}\|_F^2 - r\big)\sigma_{\mathbb{K},+}^2 + \|\boldsymbol{A}-\boldsymbol{A}^{(0)}\|_F^2\,\sigma_{\mathbb{K},\times}^2 + \sum_{l=1}^{m}\mathbb{E}\big(|\delta_l^A|^2/u^2\big)\|\boldsymbol{A}^{(l)}\|_F^2\Big). \tag{3.8}$$

Proof. By (M3) and the assumption $\mathrm{Cov}\,(\boldsymbol{X}) = \rho^2 I_n$, the total precomputation error $\hat{\boldsymbol{A}}\boldsymbol{X} - \boldsymbol{A}\boldsymbol{X} = \sum_{l=1}^m \delta_l^A \boldsymbol{A}^{(l)}\boldsymbol{X}$ has the covariance matrix

$$\sum_{l=1}^m \mathbb{E}\big(|\delta_l^A|^2\big)\boldsymbol{A}^{(l)}\mathrm{Cov}\,(\boldsymbol{X})\boldsymbol{A}^{(l)\,*} = \rho^2 \sum_{l=1}^m \mathbb{E}\big(|\delta_l^A|^2\big)\boldsymbol{A}^{(l)}\boldsymbol{A}^{(l)\,*}.$$

By (3.5) the roundoff error of the product $A_{j,k}X_k$ equals $A_{j,k}^{(l)}\Delta_k^{(l)}$, where l is the unique non–trivial summand for component (j,k) in the above decomposition of A, and we have set $\Delta_k^{(l)} := |a_l X_k|\varepsilon_{l,k}^{\times}$ for $l \geq 1$ and $\boldsymbol{\Delta}^{(l)} := 0$ in the case $l = 0$. Therefore the total roundoff errors from the multiplications is $\boldsymbol{\Delta}^{\times} = \sum_{l=1}^m \boldsymbol{A}^{(l)}\boldsymbol{\Delta}^{(l)}$. Since $\mathrm{Cov}\,(\boldsymbol{\Delta}^{(l)}) = |a_l|^2\sigma_{\mathbb{K},\times}^2\rho^2u^2\boldsymbol{I}_n$ and (M3), $\boldsymbol{\Delta}^{\times}$ has the covariance matrix $\mathrm{Cov}\,(\boldsymbol{\Delta}^{\times}) = \rho^2u^2\sum_{l=1}^m |a_l|^2\sigma_{\mathbb{K},\times}^2\boldsymbol{A}^{(l)}\boldsymbol{A}^{(l)\,*}$ and the result (i) follows. The contribution of the summations is as in Lemma 3.4. By the assumption on the intermediary results, the components of these roundoff errors are uncorrelated.

Since $\mathbb{E}\|\boldsymbol{X}\|_2^2) = n\rho^2$, $\bar{k}_n^2$ is $n^{-1}\rho^{-2}$ times the trace of $\mathrm{Cov}\,(\boldsymbol{\Delta})$. Assertion (ii) now follows from $\mathrm{tr}(\boldsymbol{B}\boldsymbol{B}^*) = \|\boldsymbol{B}\|_F^2$. ∎

3.9 Remark. Clearly, if $\boldsymbol{A}^{(l)}$ is the elementary matrix $\boldsymbol{E}_{j,k}$, we have $\boldsymbol{A}^{(l)}\boldsymbol{A}^{(l)\,*} = \boldsymbol{E}_{j,j}$. In the case of a rotation $\boldsymbol{A} = \begin{pmatrix} \cos\phi & \sin\phi \\ -\sin\phi & \cos\phi \end{pmatrix} = \cos\phi\,\boldsymbol{A}^{(1)} + \sin\phi\,\boldsymbol{A}^{(2)}$ with $\boldsymbol{A}^{(1)} = \boldsymbol{I}_2$, $\boldsymbol{A}^{(2)} = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}$ we have $\boldsymbol{A}^{(l)}\boldsymbol{A}^{(l)\,*} = \boldsymbol{I}_2$ for $l = 1, 2$. The covariance matrix of the roundoff error is therefore $\rho^2\big(\sigma_{\mathbb{K},+}^2u^2 + \sigma_{\mathbb{R},\times}^2u^2 + \mathbb{E}(|\delta^{\cos}|^2 + |\delta^{\sin}|^2)\big)\,\boldsymbol{I}_2$. If we assume that the relative error ε^{tr} in the (pre-)computation of a trigonometric function has variance $\mathbb{V}(\varepsilon^{tr}) = \sigma_{tr}^2u^2$ then this simplifies to $\rho^2\big(\sigma_{\mathbb{K},+}^2 + \sigma_{\mathbb{R},\times}^2 + \sigma_{tr}^2\big)u^2\,\boldsymbol{I}_2$. The same is true for reflections $\boldsymbol{A} = \begin{pmatrix} \cos\phi & \sin\phi \\ \sin\phi & -\cos\phi \end{pmatrix}$. If all $\boldsymbol{A}^{(j)}$ $(j \geq 1)$ in the decomposition of $\boldsymbol{A}$ are of one of these types, then the components of the roundoff errors are uncorrelated. However, in general this is not always the case.

It is an interesting consequence of Proposition 3.8 (ii) that for the multiplication with the conjugate matrix $\boldsymbol{A}^*$ the square of the norm of the roundoff error has the same expectation as for the original algorithm, although the covariance matrices are different in general. For example, this is the relation between the Cooley–Tukey and Gentleman–Sande algorithms, and between the DCT–II and DCT-III transforms. It follows from $\|\boldsymbol{B}\|_F = \|\boldsymbol{B}^*\|_F$ for all matrices $\boldsymbol{B}$.

3.10 Remark. The multiplication of a vector $\boldsymbol{X} \in \mathbb{R}^2$ with a rotation matrix $\boldsymbol{A} = \begin{pmatrix} \cos\phi & \sin\phi \\ -\sin\phi & \cos\phi \end{pmatrix}$ takes four multiplications and two summations. For $\phi \in\,]-\pi, \pi[$ it is also possible to compute $\boldsymbol{A}\boldsymbol{X}$ by only three multiplications but three additions using three *lifting steps* (see [6]), i.e. the (nonunitary) decomposition $\boldsymbol{A} = \begin{pmatrix} 1 & t \\ 0 & 1 \end{pmatrix}\begin{pmatrix} 1 & 0 \\ -s & 1 \end{pmatrix}\begin{pmatrix} 1 & t \\ 0 & 1 \end{pmatrix}$ where $s := \sin\phi$ and $t := \frac{1-\cos\phi}{\sin\phi} = \tan\frac{\phi}{2}$. If s and t are precomputed as $\hat{s} = s + \delta^{\sin}$ and $\hat{t} = t + \delta^{\tan}$, then

$$\begin{pmatrix} 1 & \hat{t} \\ 0 & 1 \end{pmatrix}\begin{pmatrix} 1 & 0 \\ -\hat{s} & 1 \end{pmatrix}\begin{pmatrix} 1 & \hat{t} \\ 0 & 1 \end{pmatrix} = \boldsymbol{A} - \begin{pmatrix} s\delta^{\tan} + t\delta^{\sin} & t^2\delta^{\sin} - 2(1-ts)\delta^{\tan} \\ \delta^{\sin} & s\delta^{\tan} + t\delta^{\sin} \end{pmatrix} + \mathcal{O}(u^2).$$

Hence the precomputation errors propagate into the final result as

$$\boldsymbol{\Delta}^{pr} = -\begin{pmatrix} s\delta^{\tan} + t\delta^{\sin} & t^2\delta^{\sin} - 2(1-ts)\delta^{\tan} \\ \delta^{\sin} & s\delta^{\tan} + t\delta^{\sin} \end{pmatrix}\begin{pmatrix} X_1 \\ X_2 \end{pmatrix}.$$

In general the components of this vector are correlated and have variances $\mathbb{V}(\Delta_1^{pr}) = (2s^2t^2 + t^4s^2 + 4(1-ts)^2t^2)\sigma_{tr}^2\rho^2u^2$ and $\mathbb{V}(\Delta_2^{pr}) = (s^2 + 2s^2t^2)\sigma_{tr}^2\rho^2u^2$. For simplification we assume as above that the relative error ε^{tr} has variance $\mathbb{V}(\varepsilon^{tr}) = \sigma_{tr}^2u^2$.

The arithmetical operations occuring in the first matrix–vector multiplication produce a roundoff error $\boldsymbol{\Delta}^{(1)}$ with $\Delta_2^{(1)} = 0$ and $\mathbb{V}(\Delta_1^{(1)}) = (t^2\sigma_{\mathbb{R},\times}^2 + (1+t^2)\sigma_{\mathbb{R},+}^2)\rho^2u^2$. It is propagated into the end result as

$$\begin{pmatrix} 1 & t \\ 0 & 1 \end{pmatrix}\begin{pmatrix} 1 & 0 \\ -s & 1 \end{pmatrix}\boldsymbol{\Delta}^{(1)} = \Delta_1^{(1)}\begin{pmatrix} \cos\phi \\ -\sin\phi \end{pmatrix}.$$

The matrix $\begin{pmatrix} 1 & 0 \\ -s & 1 \end{pmatrix}$ operates on $\begin{pmatrix} 1 & t \\ 0 & 1 \end{pmatrix}\begin{pmatrix} X_1 \\ X_2 \end{pmatrix} = \begin{pmatrix} X_1 + tX_2 \\ X_2 \end{pmatrix}$. This operation produces a roundoff error $\boldsymbol{\Delta}^{(2)}$ with $\mathbb{V}(\Delta_2^{(2)}) = (s^2(1+t^2)\sigma_{\mathbb{R},\times}^2 + \sigma_{\mathbb{R},+}^2)\rho^2u^2$ and $\Delta_1^2 = 0$. It appears in the end result as $\begin{pmatrix} 1 & t \\ 0 & 1 \end{pmatrix}\boldsymbol{\Delta}^2 = \Delta_2^2\begin{pmatrix} t \\ 1 \end{pmatrix}$. The last matrix operates on $\begin{pmatrix} 1 & 0 \\ -s & 1 \end{pmatrix}\begin{pmatrix} 1 & t \\ 0 & 1 \end{pmatrix}\begin{pmatrix} X_1 \\ X_2 \end{pmatrix} = \begin{pmatrix} X_1 + tX_2 \\ -sX_1 + (1-ts)X_2 \end{pmatrix}$. The roundoff error $\boldsymbol{\Delta}^{(3)}$ has $\Delta_2^{(3)} = 0$ and $\mathbb{V}(\Delta_1^{(3)}) = (t^2\sigma_{\mathbb{R},\times}^2 + \sigma_{\mathbb{R},+}^2)\rho^2u^2$.

In total we obtain the error vector

$$\boldsymbol{\Delta} = \boldsymbol{\Delta}^{pr} + \begin{pmatrix} 1 & t \\ 0 & 1 \end{pmatrix}\begin{pmatrix} 1 & 0 \\ -s & 1 \end{pmatrix}\boldsymbol{\Delta}^{(1)} + \begin{pmatrix} 1 & t \\ 0 & 1 \end{pmatrix}\boldsymbol{\Delta}^{(2)} + \boldsymbol{\Delta}^{(3)}$$

which has correlated components and satisfies

$$\begin{aligned}\mathbb{E}(\|\boldsymbol{\Delta}\|_2^2) &= \mathbb{V}(\Delta_1) + \mathbb{V}(\Delta_2) \\ &= \left(\left(6 + 2\cos^2\phi\right)\tan^2\frac{\phi}{2}\,\sigma_{tr}^2 + \left(3 + 2\tan^2\frac{\phi}{2}\right)\sigma_{\mathbb{R},+}^2 + 6\tan^2\frac{\phi}{2}\,\sigma_{\mathbb{R},\times}^2\right)\rho^2u^2.\end{aligned}$$

For small rotation angles ϕ, the roundoff error is less than for the classical rotation, since the precomputation and multiplication errors are smaller, but if $|\phi| \geq \frac{\pi}{3}$, the variance of the roundoff error from the lifting method is larger.

As one benefit of the use of lifting steps it is often claimed that they allow lossless reconstruction, since the inverse of $\begin{pmatrix} 1 & t \\ 0 & 1 \end{pmatrix}$ is $\begin{pmatrix} 1 & -t \\ 0 & 1 \end{pmatrix}$ and therefore the precomputation and multiplication errors in the lifting step and its inverse cancel exactly. However, apart from the innermost lifting, this property only holds if the floating point addition $x + z$ (where $z = ty$) followed by the subtraction of z yields exactly $\hat{x} = x$ again. Let the probability of this event be denoted by $p > 0$ (from simulations with standard normal data one estimates $p \approx 0.61$, but if z is small – this is the case for small angles ϕ – the probability p is close to 1). However, if this event does *not* happen, the roundoff errors from the multiplications $t \cdot x$ and $t \cdot \hat{x}$ are distinct and do not cancel. Hence in the j-th to last lifting step the probability for the cancellation of the multiplicative errors is only q^{j-1}. The precomputation errors cancel in both cases (up to terms $\mathcal{O}(u^2)$).

4. Application to DFT and DCT

A usual simplification in the average case study of the roundoff error in the fast trigonometric transforms is the assumption of independence of the roundoff errors from the butterfly operations $y_1 = x_1 + \omega x_2$, $y_2 = x_1 - \omega x_2$ (where $|\omega| = 1$). Since here the same product enters into two components, these are obviously not independent. This problem has already been noted in [4]. Also, usually for all multiplications the same roundoff error model is used, although the trivial factors $1, -1, \mathrm{i}, -\mathrm{i}$ do not induce any error.

In order to highlight the influence of these assumptions we will calculate the covariance matrix of the roundoff error *without* this simplification in the case $n = 8$. This case has the most most important applications (e.g. JPEG) and also the effects are more pronounced for small values of n. Except in Example 4.3, we always assume that the data vector $\boldsymbol{X}$ has expectation 0 and uncorrelated components of variance $\|\!|X_j|\!\|^2 = \rho^2$.

We begin with the study of the discrete Fourier transform $\hat{\boldsymbol{x}} := \boldsymbol{F}_n \boldsymbol{x}$, $(\boldsymbol{x} \in \mathbb{C}^n)$, where $\boldsymbol{F}_n := n^{-1/2}(\omega^{jk})_{j,k=0}^{n-1}$ with $\omega = \omega_n := \exp(-2\pi\mathrm{i}/n)$. In the following, $\boldsymbol{I}_n$ will always denote the identity matrix, and $\boldsymbol{J}_n$ the counteridentity matrix $(\delta_{j+k,n-1})_{j,k=0}^{n-1}$ of order n.

4.1 Example. First we study the naive algorithm for the computation of the DFT $\hat{\boldsymbol{X}} = \boldsymbol{F}_8 \boldsymbol{X}$ via the matrix–vector multiplication $(\sqrt{8}\boldsymbol{F}_8)\boldsymbol{X}$ followed by a multiplication with $1/\sqrt{8}$. It uses one function call (for $\cos\frac{2\pi}{8} = \sin\frac{2\pi}{8} = \frac{\sqrt{2}}{2}$), 56 summations (both for recursive or cascade summation) and 32 complex multiplications. Note that the matrix entries $\omega^0 = 1, \omega^2 = -\mathrm{i}, \omega^4 = -1$ and $\omega^6 = \mathrm{i}$ give trivial multiplications. Furthermore, if we exploit the special form of $\omega = \frac{\sqrt{2}}{2}(1 - \mathrm{i})$, only four complex products have to be computed.

Let $\boldsymbol{A}^{(0)}, \boldsymbol{A}^{(1)} \in \mathbb{C}^{8\times 8}$ be defined by

$$A^{(0)}_{j,k} := \begin{cases} (-\mathrm{i})^{jk/2} & \text{for } jk \text{ even}, \\ 0 & \text{for } jk \text{ odd}, \end{cases} \qquad A^{(1)}_{j,k} := \begin{cases} 0 & \text{for } jk \text{ even}, \\ (-\mathrm{i})^{(jk-1)/2} & \text{for } jk \text{ odd}. \end{cases}$$

Then we have

$$\sqrt{8}\boldsymbol{F}_8 = \boldsymbol{A}^{(0)} + \omega \boldsymbol{A}^{(1)}$$

and we can apply Proposition 3.8. If the direct call results in $\mathrm{fl}\big(\frac{\sqrt{2}}{2}\big) = \frac{\sqrt{2}}{2} + \delta^{\surd}$, then the computed value for $a_1 := \omega$ is $\hat{\omega} = \omega(1 + \sqrt{2}\delta^{\surd})$, which has variance $2\sigma^2_{\surd} = 2\mathbb{V}(\delta^{\surd}) = u^2/6$. Here we assume that $\delta^{\surd}$ is uniformly distributed on an interval of length u. Hence in (3.6) we have $\mathbb{V}(\delta_1^A) + |a_1|^2\sigma^2_{\mathbb{K},\times} = 2\sigma^2_{\surd} + \sigma^2_{\mathbb{C},\times}$ and by a simple calculation we obtain $\boldsymbol{A}^{(1)}\boldsymbol{A}^{(1)^*} = 4\boldsymbol{K}$ with

$$\boldsymbol{K} := \boldsymbol{E}_{1,1} + \boldsymbol{E}_{3,3} + \boldsymbol{E}_{5,5} + \boldsymbol{E}_{7,7} - \boldsymbol{E}_{1,5} - \boldsymbol{E}_{5,1} - \boldsymbol{E}_{3,7} - \boldsymbol{E}_{7,3}\,.$$

In the case of recursive summation the variance of each component of $\boldsymbol{\Sigma}^+$ is $\Sigma^+_{j,j} = \big(7 + \sum_{k=1}^{7}(8-k)\big)\sigma^2_{\mathbb{C},+}\rho^2 = 35\sigma^2_{\mathbb{C},+}\rho^2$ for $j = 1, \ldots, 8$. In the case of cascade summation this has to be replaced by $\Sigma^+_{j,j} = 3 \cdot 8\sigma^2_{\mathbb{C},+}\rho^2 = 24\sigma^2_{\mathbb{C},+}\rho^2$.

Taking into account the final multiplication with $1/\sqrt{8}$, we obtain for the total

roundoff error vector $\boldsymbol{\Delta} = \mathrm{fl}(\hat{\boldsymbol{F}}_8\boldsymbol{X}) - \boldsymbol{F}_8\boldsymbol{X}$ the covariance matrix

$$\mathrm{Cov}\,(\boldsymbol{\Delta}) = \begin{pmatrix} \alpha & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & \alpha+\beta & 0 & 0 & 0 & -\beta & 0 & 0 \\ 0 & 0 & \alpha & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & \alpha+\beta & 0 & 0 & 0 & -\beta \\ 0 & 0 & 0 & 0 & \alpha & 0 & 0 & 0 \\ 0 & -\beta & 0 & 0 & 0 & \alpha+\beta & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & \alpha & 0 \\ 0 & 0 & 0 & -\beta & 0 & 0 & 0 & \alpha+\beta \end{pmatrix} \rho^2 u^2 \tag{4.2}$$

where $\beta = \frac{4}{8}(\mathbb{V}(\delta_1^A)+|a_1|^2\sigma^2_{\mathbb{K},\times}) = \sigma^2_{\surd}+\frac{1}{2}\sigma^2_{\mathbb{C},\times}$, and $\alpha = \frac{35}{8}\sigma^2_{\mathbb{C},+}$ for recursive summation resp. $\alpha = 3\sigma^2_{\mathbb{C},+}$ for cascade summation. In particular, since $\mathbb{E}(\|\boldsymbol{X}\|_2^2) = 8\rho^2$,

$$\mathbb{E}\big(\|\boldsymbol{\Delta}\|_2^2\big) = (8\alpha+4\beta)\rho^2u^2 = \Big(\alpha+\frac{1}{2}\beta\Big)\,\mathbb{E}\big(\|\boldsymbol{X}\|_2^2\big)u^2.$$

4.2 Remark. The precomputed matrix entries are considered random in the computations above, uniformly distributed between the two next machine numbers. But if the arithmetic of the computer is fixed, unlike the effects of the randomness of the data, the roundoff error δ^ω of a precomputed constant ω remains the same every instance the algorithm is called. Therefore, in simulations we obtain results which are conditioned on these specific values. This explains the strong irregularities in the mean error of an FFT for different orders n (see for example [7], Fig. 23.1 or [13], Fig. 3) which often are much larger than the standard deviation of the errors.

We can therefore obtain a model better fitting the roundoff errors, if we replace $\sigma^2_\omega = \mathbb{E}(|\delta^\omega|^2)$ with the square of the *exact* error (which, for example in the calculation of the constant $\frac{1}{2}\sqrt{2}$ above, is $0.190u^2$ for 53-bit mantissa but only $0.041u^2$ for 24-bit mantissa, compared with the expected value $\sigma^2_\omega = \frac{1}{12}u^2$). Proposition 3.8 is still applicable under this situation. Naturally, this approach is only possible if we are able to compute the entries to higher precision, and it is cumbersome when the matrix has a large number of different precomputed entries. If we cannot apply this more precise model, we have to use $\mathbb{E}(|\delta|^2) = 2^{-2k}u^2/12$ for $2^{-k-1} < |\omega| < 2^{-k}$ ($k \in \mathbb{Z}$), which follows from the fact that in this interval the distance between two consecutive machine numbers is $2^{-k}u$.

4.3 Remark. Let us study for a moment the influence of non–zero expectation of the data vector $\boldsymbol{X}$ in the case of cascade summation. If $\mathbb{E}(\boldsymbol{X}) = \boldsymbol{\mu} = (\mu_j)_{j=0}^7$ and $\rho^2 := \mathbb{V}(X_j)$, then $\|\!|X_j|\!\|^2 = \rho^2+|\mu_j|^2$. It follows from Lemma 3.6 that we now have

$$\begin{aligned}\Sigma^+_{k,k} &= 3\sum_{j=0}^{7}(\rho^2+|\mu_j|^2)\,\sigma^2_{\mathbb{C},+} + \Big(\Big|\sum_{j=0}^{7}\omega^{jk}\mu_j\Big|^2 - \sum_{j=0}^{7}|\mu_j|^2\Big)\sigma^2_{\mathbb{C},+} \\ &= \big(24\rho^2+2\|\boldsymbol{\mu}\|^2+8|\hat{\boldsymbol{\mu}}(k)|^2\big)\sigma^2_{\mathbb{C},+}\end{aligned}$$

with $\hat{\boldsymbol{\mu}} = \mathbb{E}(\hat{\boldsymbol{X}})$. Hence in (4.2) we have to replace the α in the k-th diagonal entry by $\alpha_k := \big(3+\frac{1}{4\rho^2}\|\boldsymbol{\mu}\|^2+\frac{1}{\rho^2}|\hat{\boldsymbol{\mu}}(k)|^2\big)\sigma^2_{\mathbb{C},+}$. Likewise, β has to be replaced by $\beta' := \big(1+\frac{1}{4\rho^2}(|\mu_1|^2+|\mu_3|^2+|\mu_5|^2+|\mu_7|^2)\big)\big(\sigma^2_{\surd}+\frac{1}{2}\sigma^2_{\mathbb{C},\times}\big)$. Furthermore, since the matrix

$\big(\mathbb{E}(X_j\,\overline{X_k})\big)_{j,k=1,3,5,7}$ is no longer a multiple of the identity matrix, additional terms outside the diagonal appear, and we obtain

$$\mathrm{Cov}\,(\boldsymbol{\Delta}) = \begin{pmatrix} \alpha_0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & \alpha_1+\beta' & 0 & \gamma\mathrm{i} & 0 & -\beta' & 0 & -\gamma\mathrm{i} \\ 0 & 0 & \alpha_2 & 0 & 0 & 0 & 0 & 0 \\ 0 & -\gamma\mathrm{i} & 0 & \alpha_3+\beta' & 0 & \gamma\mathrm{i} & 0 & -\beta' \\ 0 & 0 & 0 & 0 & \alpha_4 & 0 & 0 & 0 \\ 0 & -\beta' & 0 & -\gamma\mathrm{i} & 0 & \alpha_5+\beta' & 0 & \gamma\mathrm{i} \\ 0 & 0 & 0 & 0 & 0 & 0 & \alpha_6 & 0 \\ 0 & \gamma\mathrm{i} & 0 & -\beta' & 0 & -\gamma\mathrm{i} & 0 & \alpha_7+\beta' \end{pmatrix} \rho^2u^2$$

with $\gamma := \frac{1}{4\rho^2}\big(|\mu_1|^2-|\mu_3|^2+|\mu_5|^2-|\mu_7|^2\big)\big(\sigma^2_{\diagup\!\!\diagdown}+\frac{1}{2}\sigma^2_{\mathbb{C},\times}\big)$.

The FFT is based on the factorization of the Fourier matrix

$$\boldsymbol{F}_n = n^{-1/2}\boldsymbol{A}_t\ldots\boldsymbol{A}_1\boldsymbol{P}_n \quad (n=2^t),$$

where $\boldsymbol{P}_n$ is the bit–reversal permutation matrix and each $\boldsymbol{A}_j$ has exactly two non–zero entries in each row, one of them $= 1$. In our special case $n = 2^3$,

$$\boldsymbol{A}_3 := \begin{pmatrix} \boldsymbol{I}_4 & \boldsymbol{D}_4 \\ \boldsymbol{I}_4 & -\boldsymbol{D}_4 \end{pmatrix} \quad \text{with } \boldsymbol{D}_4 := \mathrm{diag}(1,\omega,\omega^2,\omega^3),$$

$$\boldsymbol{A}_2 := \mathrm{diag}\left(\begin{pmatrix} \boldsymbol{I}_2 & \boldsymbol{D}_2 \\ \boldsymbol{I}_2 & -\boldsymbol{D}_2 \end{pmatrix}, \begin{pmatrix} \boldsymbol{I}_2 & \boldsymbol{D}_2 \\ \boldsymbol{I}_2 & -\boldsymbol{D}_2 \end{pmatrix}\right) \quad \text{with } \boldsymbol{D}_2 := \mathrm{diag}(1,-\mathrm{i}),$$

and $\boldsymbol{A}_1$ is block–diagonal with four blocks $\begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}$.

4.4 Example. In the computation of the FFT $\hat{\boldsymbol{X}} = 8^{-1/2}\boldsymbol{A}_3\boldsymbol{A}_2\boldsymbol{A}_1\boldsymbol{P}_8\boldsymbol{X}$, one direct call, four complex multiplications and 24 additions are used, since only A_3 contains nontrivial factors ω and $\omega^3 = -\mathrm{i}\omega$. As in the case of the naive algorithm we do not take into account the error from the final multiplication by $8^{-1/2}$.

The butterfly operations in the multiplication with $\boldsymbol{A}_1$ and $\boldsymbol{A}_2$ give each rise to independent roundoff errors $\boldsymbol{\Delta}_1^+$ and $\boldsymbol{\Delta}_2^+$ of variance $2\sigma^2_{\mathbb{C},+}\rho^2u^2$ resp. $4\sigma^2_{\mathbb{C},+}\rho^2u^2$ in every component – no multiplications are necessary. Then $\boldsymbol{X}^{(2)} := \boldsymbol{A}_2\boldsymbol{A}_1\boldsymbol{P}_8\boldsymbol{X}$ has uncorrelated components of variance $4\rho^2$.

Let $\boldsymbol{A}^{(0)} := \begin{pmatrix} \boldsymbol{I}_4 & \boldsymbol{D}'_4 \\ \boldsymbol{I}_4 & -\boldsymbol{D}'_4 \end{pmatrix}$ with $\boldsymbol{D}'_4 := \mathrm{diag}(1,0,-\mathrm{i},0)$, and $\boldsymbol{A}^{(1)} := \boldsymbol{E}_{1,5}-\mathrm{i}\boldsymbol{E}_{3,7}-\boldsymbol{E}_{5,5}+\mathrm{i}\boldsymbol{E}_{7,7}$. Since

$$\boldsymbol{A}_3 = \boldsymbol{A}^{(0)} + \omega\boldsymbol{A}^{(1)}$$

the roundoff error $\boldsymbol{\Delta}_3$ from the multiplication $\boldsymbol{A}_3\boldsymbol{X}^{(2)}$ has the covariance matrix $\mathrm{Cov}\,(\boldsymbol{\Delta}_3) = 8\rho^2\sigma^2_{\mathbb{C},+}\boldsymbol{I}_8u^2 + 4\rho^2(2\sigma^2_{\diagup\!\!\diagdown}+\sigma^2_{\mathbb{C},\times})\boldsymbol{K}u^2$ by Proposition 3.8, where $\boldsymbol{K}$ and $\sigma^2_{\diagup\!\!\diagdown}$ have the same meaning as in Example 4.1.

Using the fact that $2^{-1/2}\boldsymbol{A}_j$ is an orthogonal matrix $(j = 1,2,3)$, the total roundoff error $\boldsymbol{\Delta} = \frac{1}{\sqrt{8}}\big(\boldsymbol{\Delta}_3 + \boldsymbol{A}_3\boldsymbol{\Delta}_2 + \boldsymbol{A}_3\boldsymbol{A}_2\boldsymbol{\Delta}_1\big)$ has the covariance matrix

$$\begin{aligned}\mathrm{Cov}\,(\boldsymbol{\Delta}) &= \frac{1}{8}\big(\mathrm{Cov}\,(\boldsymbol{\Delta}_3) + \boldsymbol{A}_3\mathrm{Cov}\,(\boldsymbol{\Delta}_2)\boldsymbol{A}_3^* + \boldsymbol{A}_3\boldsymbol{A}_2\mathrm{Cov}\,(\boldsymbol{\Delta}_1)\boldsymbol{A}_2^*\boldsymbol{A}_3^*\big) \\ &= 3\sigma^2_{\mathbb{C},+}\rho^2\boldsymbol{I}_8u^2 + \rho^2(\sigma^2_{\diagup\!\!\diagdown}+\frac{1}{2}\sigma^2_{\mathbb{C},\times})\boldsymbol{K}u^2,\end{aligned}$$

which is the same matrix as (4.2) with $\beta = \sigma_{\checkmark}^2 + \frac{1}{2}\sigma_{\mathbb{C},\times}^2$ and $\alpha = 3\sigma_{\mathbb{C},+}^2$. Again,

$$\mathbb{E}\big(\|\boldsymbol{\Delta}\|_2^2\big) = \Big(\alpha + \frac{1}{2}\beta\Big)\,\mathbb{E}\big(\|\boldsymbol{X}\|_2^2\big)$$

with the same α and β as in the naive algorithm with cascade summation.

4.5 Example. With the Gentleman–Sande algorithm, the FFT is computed via $\hat{\boldsymbol{X}} = 8^{-1/2}\boldsymbol{P}_8\boldsymbol{A}_1^T\boldsymbol{A}_2^T\boldsymbol{A}_3^T\boldsymbol{X}$. Since $\boldsymbol{A}_3^{\mathrm{T}} = \boldsymbol{A}^{(0)} + \omega\boldsymbol{A}^{(1)}$ with $\boldsymbol{A}^{(0)} := \begin{pmatrix} \boldsymbol{I}_4 & \boldsymbol{I}_4 \\ \boldsymbol{D}_4'' & -\boldsymbol{D}_4'' \end{pmatrix}$, $\boldsymbol{D}_4'' := \mathrm{diag}(0,1,0,-\mathrm{i})$ and $\boldsymbol{A}^{(1)} := -\boldsymbol{E}_{1,5} + \mathrm{i}\boldsymbol{E}_{3,7} + \boldsymbol{E}_{5,1} - \mathrm{i}\boldsymbol{E}_{7,3}$, it follows from Proposition 3.8 that the roundoff error $\boldsymbol{\Delta}_3$ from the product $\boldsymbol{A}_3^{\mathrm{T}}\boldsymbol{X}$ has the covariance matrix $\mathrm{Cov}\,(\boldsymbol{\Delta}_3) = 2\rho^2\sigma_{\mathbb{C},+}^2\boldsymbol{I}_8u^2 + \rho^2(2\sigma_{\checkmark}^2 + \sigma_{\mathbb{C},\times}^2)\boldsymbol{K}'u^2$ with $\boldsymbol{K}' := \boldsymbol{A}^{(1)}\boldsymbol{A}^{(1)\,*} = 2\boldsymbol{E}_{5,5}+2\boldsymbol{E}_{7,7}$. Since later no proper multiplications occur, and since the components of $\boldsymbol{A}_3^{\mathrm{T}}\boldsymbol{X}$ and $\boldsymbol{A}_2^{\mathrm{T}}\boldsymbol{A}_3^{\mathrm{T}}\boldsymbol{X}$ have variance $2\rho^2$ resp. $4\rho^2$, the roundoff errors $\boldsymbol{\Delta}_2$ and $\boldsymbol{\Delta}_1$ from the product with $\boldsymbol{A}_2^{\mathrm{T}}$ resp. $\boldsymbol{A}_1^{\mathrm{T}}$ have covariance matrices $\mathrm{Cov}\,(\boldsymbol{\Delta}_2) = 4\rho^2\sigma_{\mathbb{C},+}^2\boldsymbol{I}_8u^2$ resp. $\mathrm{Cov}\,(\boldsymbol{\Delta}_1) = 8\rho^2\sigma_{\mathbb{C},+}^2\boldsymbol{I}_8u^2$.

In total, $\boldsymbol{\Delta} := \frac{1}{\sqrt{8}}\boldsymbol{P}_8\big(\boldsymbol{\Delta}_1 + \boldsymbol{A}_1^{\mathrm{T}}\boldsymbol{\Delta}_2 + \boldsymbol{A}_1^{\mathrm{T}}\boldsymbol{A}_2^{\mathrm{T}}\boldsymbol{\Delta}_3\big)$ has the covariance matrix

$$\mathrm{Cov}\,(\boldsymbol{\Delta}) = \mathrm{diag}(\alpha,\alpha,\alpha,\alpha,\alpha,\alpha+2\beta,\alpha,\alpha+2\beta)u^2$$

with $\alpha = 3\sigma_{\mathbb{C},+}^2$ and $\beta = \sigma_{\checkmark}^2 + \frac{1}{2}\sigma_{\mathbb{C},\times}^2$. We have used the identities $\boldsymbol{A}_2\boldsymbol{K}' = \boldsymbol{K}'\boldsymbol{A}_2$, $\boldsymbol{A}_1\boldsymbol{K}' = \boldsymbol{K}'\boldsymbol{A}_1$ and $\boldsymbol{P}\boldsymbol{K}' = \boldsymbol{K}'\boldsymbol{P}$. As in the example above, we obtain

$$\mathbb{E}\big(\|\boldsymbol{\Delta}\|_2^2\big) = \Big(\alpha + \frac{1}{2}\beta\Big)\,\mathbb{E}\big(\|\boldsymbol{X}\|_2^2\big)u^2.$$

We will now study the discrete cosine transformations (DCT) of type II and III. The DCT-II $\hat{\boldsymbol{x}} := \boldsymbol{C}_n^{\mathrm{II}}\boldsymbol{x}$ ($\boldsymbol{x} \in \mathbb{R}^n$) uses the cosine-II matrix

$$\boldsymbol{C}_n^{\mathrm{II}} = \frac{1}{\sqrt{n}}\Big(\eta_j \cos\Big(j(2k+1)\frac{\pi}{2n}\Big)\Big)_{j,k=0}^{n-1} \quad \text{with } \eta_j = \begin{cases} 1 & \text{for } j = 0, \\ \sqrt{2} & \text{for } j = 1,\dots,n-1. \end{cases}$$

We will again treat the case $n = 8$ only, which is of special interest in the JPEG algorithm of image compression.

4.6 Example. In the naive approach of the matrix–vector product, 64 multiplications and 56 summations are used. By examining the entries of $\boldsymbol{C}_8^{\mathrm{II}}$ we see that

$$\boldsymbol{C}_8^{\mathrm{II}} = \sum_{j=1}^{7} c_j\boldsymbol{C}^{(j)}$$

with $c_1 = \frac{1}{2\sqrt{2}} = \frac{1}{2}\cos\frac{\pi}{4}$, $c_2 = \frac{1}{2}\cos\frac{\pi}{8}$, $c_3 = \frac{1}{2}\cos\frac{3\pi}{8}$, $c_4 = \frac{1}{2}\cos\frac{\pi}{16}$, $c_5 = \frac{1}{2}\cos\frac{3\pi}{16}$, $c_6 = \frac{1}{2}\cos\frac{5\pi}{16}$, and $c_7 = \frac{1}{2}\cos\frac{7\pi}{16}$ occuring at

$$\begin{aligned}
\boldsymbol{C}^{(1)} &= \textstyle\sum_{k=0}^{7}\boldsymbol{E}_{0,k} + \sum_{k=0,3,4,7}\boldsymbol{E}_{4,k} - \sum_{k=1,2,5,6}\boldsymbol{E}_{4,k}, \\
\boldsymbol{C}^{(2)} &= \boldsymbol{E}_{2,0} - \boldsymbol{E}_{2,3} - \boldsymbol{E}_{2,4} + \boldsymbol{E}_{2,7} - \boldsymbol{E}_{6,1} + \boldsymbol{E}_{6,2} + \boldsymbol{E}_{6,5} - \boldsymbol{E}_{6,6}, \\
\boldsymbol{C}^{(3)} &= \boldsymbol{E}_{2,1} - \boldsymbol{E}_{2,2} - \boldsymbol{E}_{2,5} + \boldsymbol{E}_{2,6} + \boldsymbol{E}_{6,0} - \boldsymbol{E}_{6,3} - \boldsymbol{E}_{6,4} + \boldsymbol{E}_{6,7}, \\
\boldsymbol{C}^{(4)} &= \boldsymbol{E}_{1,0} - \boldsymbol{E}_{1,7} - \boldsymbol{E}_{3,2} + \boldsymbol{E}_{3,5} - \boldsymbol{E}_{5,1} + \boldsymbol{E}_{5,6} - \boldsymbol{E}_{7,3} + \boldsymbol{E}_{7,4}, \\
\boldsymbol{C}^{(5)} &= \boldsymbol{E}_{1,1} - \boldsymbol{E}_{1,6} + \boldsymbol{E}_{3,0} - \boldsymbol{E}_{3,7} + \boldsymbol{E}_{5,3} - \boldsymbol{E}_{5,4} + \boldsymbol{E}_{7,2} - \boldsymbol{E}_{7,5}, \\
\boldsymbol{C}^{(6)} &= \boldsymbol{E}_{1,2} - \boldsymbol{E}_{1,5} - \boldsymbol{E}_{3,3} + \boldsymbol{E}_{3,4} + \boldsymbol{E}_{5,0} - \boldsymbol{E}_{5,7} - \boldsymbol{E}_{7,1} + \boldsymbol{E}_{7,6}, \\
\boldsymbol{C}^{(7)} &= \boldsymbol{E}_{1,3} - \boldsymbol{E}_{1,4} - \boldsymbol{E}_{3,1} + \boldsymbol{E}_{3,6} + \boldsymbol{E}_{5,2} - \boldsymbol{E}_{5,5} + \boldsymbol{E}_{7,0} - \boldsymbol{E}_{7,7}\,.
\end{aligned}$$

Since $c_j \in \,]\frac{1}{4},\frac{1}{2}[$ for $j = 1,2,4,5,6$, $c_3 \in \,]\frac{1}{8},\frac{1}{4}[$ and $c_7 \in \,]\frac{1}{16},\frac{1}{8}[$, and assuming that the exact values are uniformly distributed between the two nearest machine numbers, we have $\mathbb{E}(|\delta_j^C|^2) = \frac{u^2}{48}$ for $j = 1,2,4,5,6$, $\mathbb{E}(|\delta_3^C|^2) = \frac{u^2}{192}$ and $\mathbb{E}(|\delta_7^C|^2) = \frac{u^2}{768}$. As in the case of $\delta^{\surd}$ in Remark 4.2, these values should be replaced by the squares of the exact roundoff errors if they are known. By simple calculations we obtain $\boldsymbol{C}^{(1)}\boldsymbol{C}^{(1)^{\mathrm{T}}} = 8\boldsymbol{E}_{0,0} + 8\boldsymbol{E}_{4,4}$, $\boldsymbol{C}^{(2)}\boldsymbol{C}^{(2)^{\mathrm{T}}} = \boldsymbol{C}^{(3)}\boldsymbol{C}^{(3)^{\mathrm{T}}} = 4\boldsymbol{E}_{2,2} + 4\boldsymbol{E}_{6,6}$ and $\boldsymbol{C}^{(j)}\boldsymbol{C}^{(j)^{\mathrm{T}}} = 2\boldsymbol{E}_{1,1} + 2\boldsymbol{E}_{3,3} + 2\boldsymbol{E}_{5,5} + 2\boldsymbol{E}_{7,7}$ for $j = 4,5,6,7$. Using $8c_1^2 = 1$, $4(c_2^2 + c_3^2) = 1$ and $2(c_4^2 + c_5^2 + c_6^2 + c_7^2) = 1$, it follows from Proposition 3.8 that the roundoff error vector $\boldsymbol{\Delta}$ has covariance

$$\operatorname{Cov}(\boldsymbol{\Delta}) = (3\sigma_{\mathbb{R},+}^2 + \sigma_{\mathbb{R},\times}^2)\rho^2u^2\boldsymbol{I}_8 + \operatorname{diag}(64,49,40,49,64,49,40,49)\frac{\rho^2u^2}{384},$$

if cascade summation is used, since $\Sigma_{j,j}^{+} = \frac{3}{8}\sum_{k=0}^{7}\eta_j^2\cos^2(j\pi\frac{2k+1}{16})\rho^2u^2\sigma_{\mathbb{R},+}^2 = 3\rho^2\sigma_{\mathbb{R},+}^2u^2$. In the norm we have

$$\mathbb{E}(\|\boldsymbol{\Delta}\|_2^2) = \Big(3\sigma_{\mathbb{R},+}^2 + \sigma_{\mathbb{R},\times}^2 + \frac{101}{768}\Big)\mathbb{E}(\|\boldsymbol{X}\|_2^2)u^2.$$

4.7 Example. Following Loeffler et al. [9], for $n = 8$ the fast DCT-II is based on the formula

$$\boldsymbol{C}_8^{\mathrm{II}} = \frac{1}{\sqrt{8}}\boldsymbol{P}\boldsymbol{L}^{(4)}\boldsymbol{L}^{(3)}\boldsymbol{L}^{(2)}\boldsymbol{L}^{(1)}$$

with $\boldsymbol{L}^{(1)} := \begin{pmatrix}\boldsymbol{I}_4 & \boldsymbol{J}_4\\ \boldsymbol{J}_4 & -\boldsymbol{I}_4\end{pmatrix}$, $\boldsymbol{L}^{(2)} := \operatorname{diag}(\boldsymbol{I}_4, \boldsymbol{R}_4)$ (where $\boldsymbol{R}_4$ is the product of the Givens rotations $\boldsymbol{G}(2,3,\frac{\pi}{16})$ and $\boldsymbol{G}(1,4,\frac{3\pi}{16})$, see [7], p. 372), $\boldsymbol{L}^{(3)}$ is block diagonal consisting of four butterfly matrices $\begin{pmatrix}1 & 1\\ 1 & -1\end{pmatrix}$, $\boldsymbol{L}^{(4)} := \operatorname{diag}(\boldsymbol{R}_4', \boldsymbol{B}_4')$ (where $\boldsymbol{R}_4'$ is block diagonal with a butterfly matrix and $\sqrt{2}\,\boldsymbol{G}(3,4,\frac{3\pi}{8})$, and $\boldsymbol{B}_4'$ consists of a butterfly in the first and last components and $\sqrt{2}\boldsymbol{I}_2$ in the middle components) and $\boldsymbol{P}$ is the permutation matrix with $\boldsymbol{P}\boldsymbol{a} = (a_0,a_7,a_2,a_5,a_1,a_6,a_3,a_4)^{\mathrm{T}}$. It uses 14 multiplications and 26 additions (not counting the final multiplication by $8^{-1/2}$). If the rotations are replaced by lifting steps (see Remark 3.10), it takes 11 multiplications and 29 additions. The roundoff errors for both versions are approximately of the same magnitude. In the following we study the first method:

In the first step no multiplications occur and therefore the computation of $\boldsymbol{X}^{(1)} = \boldsymbol{L}^{(1)}\boldsymbol{X}$ produces a roundoff error $\boldsymbol{\Delta}^{(1)}$ with covariance matrix $\boldsymbol{K}^{(1)} = 2\rho^2\sigma_{\mathbb{R},+}^2u^2\boldsymbol{I}_8$. $\boldsymbol{X}^{(1)}$ has uncorrelated component of variance $2\rho^2$. It follows from Remark 3.9 that the covariance matrix of the error $\boldsymbol{\Delta}^{(2)}$ from the computation of $\boldsymbol{X}^{(2)} = \boldsymbol{L}^{(2)}\boldsymbol{X}^{(1)}$ is $\boldsymbol{K}^{(2)} = 2\rho^2\operatorname{diag}(0,0,0,0,k_3,k_1,k_1,k_3)u^2$ with

$$k_j := \sigma_{\mathbb{R},+}^2 + \sigma_{\mathbb{R},\times}^2 + \frac{1}{u^2}\mathbb{E}\big(|\delta^{\cos(j\pi/16)}|^2 + |\delta^{\sin(j\pi/16)}|^2\big) = \begin{cases}\sigma_{\mathbb{R},+}^2 + \sigma_{\mathbb{R},\times}^2 + \frac{17}{192} & j = 1,\\ \sigma_{\mathbb{R},+}^2 + \sigma_{\mathbb{R},\times}^2 + \frac{1}{6} & j = 3.\end{cases}$$

For the precomputation errors we have used that $\frac{1}{8} < \sin\frac{\pi}{16} < \frac{1}{4}$ and $\frac{1}{2} < \sin\frac{3\pi}{16}, \cos\frac{3\pi}{16}, \cos\frac{\pi}{16} < 1$.

From the third step $\boldsymbol{X}^{(3)} = \boldsymbol{L}^{(3)}\boldsymbol{X}^{(2)}$ we obtain a roundoff error $\boldsymbol{\Delta}^{(3)}$ of covariance matrix $\boldsymbol{K}^{(3)} = 4\rho^2\sigma^2_{\mathbb{R},+}u^2\boldsymbol{I}_8$. Finally, since each component $X_j^{(3)}$ has variance $4\rho^2$, in the last non trivial product $\boldsymbol{X}^{(4)} = \boldsymbol{L}^{(4)}\boldsymbol{X}^{(3)}$ a roundoff error $\boldsymbol{\Delta}^{(4)}$ with covariance matrix $\boldsymbol{K}^{(4)} = 4\rho^2\,\mathrm{diag}(2\sigma^2_{\mathbb{R},+}, 2\sigma^2_{\mathbb{R},+}, k_0, k_0, 2\sigma^2_{\mathbb{R},+}, k_2, k_2, 2\sigma^2_{\mathbb{R},+})u^2$ arises, where $k_0 = 2\sigma^2_{\mathbb{R},+}+2\sigma^2_{\mathbb{R},\times}+u^{-2}\mathbb{E}(|\delta^{\sqrt{2}\cos(3\pi/8)}|^2+|\delta^{\sqrt{2}\sin(3\pi/8)}|^2) = 2\sigma^2_{\mathbb{R},+}+2\sigma^2_{\mathbb{R},\times}+5/12$ (since $\frac{1}{2} < \sqrt{2}\cos\frac{3\pi}{8} < 1 < \sqrt{2}\sin\frac{3\pi}{8} < 2$) and $k_2 = 2\sigma^2_{\mathbb{R},\times} + u^{-2}\mathbb{E}(|\delta^{\sqrt{\ }}|^2) = 2\sigma^2_{\mathbb{R},\times} + 1/3$.

We will now study the covariance matrix of the total roundoff error

$$\boldsymbol{\Delta} := 8^{-1/2}\boldsymbol{P}\Big(\boldsymbol{\Delta}^{(4)} + \boldsymbol{L}^{(4)}\boldsymbol{\Delta}^{(3)} + \boldsymbol{L}^{(4)}\boldsymbol{L}^{(3)}\boldsymbol{\Delta}^{(2)} + \boldsymbol{L}^{(4)}\boldsymbol{L}^{(3)}\boldsymbol{L}^{(2)}\boldsymbol{\Delta}^{(1)}\Big).$$

A simple calculation shows that $8^{-1/2}\boldsymbol{P}\boldsymbol{L}^{(4)}\boldsymbol{L}^{(3)}\boldsymbol{\Delta}^{(2)}$ has the covariance matrix

$$\frac{1}{8}\boldsymbol{P}\boldsymbol{L}^{(4)}\boldsymbol{L}^{(3)}\boldsymbol{K}^{(2)}\boldsymbol{L}^{(3)^{\mathrm{T}}}\boldsymbol{L}^{(4)^{\mathrm{T}}}\boldsymbol{P}^{\mathrm{T}} = \Big(\frac{k_1+k_3}{2}\,\mathrm{diag}(0,1,0,1,0,1,0,1)+\frac{k_3-k_1}{2\sqrt{2}}\boldsymbol{M}_8\Big)\rho^2u^2$$

with $\boldsymbol{M}_8 := \boldsymbol{E}_{1,3}+\boldsymbol{E}_{3,1}+\boldsymbol{E}_{1,5}+\boldsymbol{E}_{5,1}+\boldsymbol{E}_{3,7}+\boldsymbol{E}_{7,3}-\boldsymbol{E}_{5,7}-\boldsymbol{E}_{7,5}$. Since $\boldsymbol{K}^{(1)}$ and $\boldsymbol{K}^{(3)}$ are multiples of $\boldsymbol{I}_8$, they pose no problem. Hence the covariance matrix of $\boldsymbol{\Delta}$ has the diagonal entries

$$\mathbb{V}(\Delta_j) = \begin{cases} 3\sigma^2_{\mathbb{R},+}\rho^2u^2 & j=0,4,\\ \big(4\sigma^2_{\mathbb{R},+}+\sigma^2_{\mathbb{R},\times}+\frac{49}{384}\big)\rho^2u^2 & j=1,7,\\ \big(3\sigma^2_{\mathbb{R},+}+\sigma^2_{\mathbb{R},\times}+\frac{5}{24}\big)\rho^2u^2 & j=2,6,\\ \big(3\sigma^2_{\mathbb{R},+}+2\sigma^2_{\mathbb{R},\times}+\frac{113}{384}\big)\rho^2u^2 & j=3,5,\end{cases}$$

plus the following non–zero entries outside the diagonal:

$$\mathrm{Cov}\,(\boldsymbol{\Delta})_{j,k} = \begin{cases} \frac{5\sqrt{2}}{256}u^2\rho^2 & (j,k)=(1,3),(1,5),(3,1),(3,7),(5,1),(7,3),\\ -\frac{5\sqrt{2}}{256}u^2\rho^2 & (j,k)=(5,7),(7,5).\end{cases}$$

In the norm, this sums up to

$$\mathbb{E}(\|\boldsymbol{\Delta}\|_2^2) = \sum_{j=0}^{7}\mathbb{V}(\Delta_j) = \Big(\frac{13}{4}\sigma^2_{\mathbb{R},+}+\sigma^2_{\mathbb{R},\times}+\frac{121}{768}\Big)\,\mathbb{E}(\|\boldsymbol{X}\|_2^2)u^2,$$

which is slightly larger than for the naive Algorithm 4.6. Other algorithms for the computation of the DCT-II in the case $n = 8$ have been described by Chen et al. [5] and Schreiber (see [12]). Using a larger number of multiplications they both satisfy

$$\mathbb{E}(\|\boldsymbol{\Delta}\|_2^2) = \Big(\frac{13}{4}\sigma^2_{\mathbb{R},+}+\frac{5}{4}\sigma^2_{\mathbb{R},\times}+\frac{133}{768}\Big)\,\mathbb{E}(\|\boldsymbol{X}\|_2^2)u^2$$

which is slightly larger than the roundoff error of the algorithm by Loeffler et al.

The DCT-III $\hat{\boldsymbol{x}} := \boldsymbol{C}_n^{\mathrm{III}}\boldsymbol{x}$ ($\boldsymbol{x}\in\mathbb{R}^n$) uses the transposed matrix $\boldsymbol{C}_n^{\mathrm{III}} := \big(\boldsymbol{C}_n^{\mathrm{II}}\big)^{\mathrm{T}}$. By orthogonality of $\boldsymbol{C}_8^{\mathrm{II}}$ it is the inverse of the cosine-II matrix.

4.8 Example. In order to study the naive algorithm of matrix–vector multiplication in the case $n = 8$, we now apply the decomposition $\boldsymbol{C}_8^{\mathrm{III}} = \big(\boldsymbol{C}_n^{\mathrm{II}}\big)^{\mathrm{T}} = \sum_{j=1}^{7} c_j \boldsymbol{C}^{(j)^{\mathrm{T}}}$, using the notations from Example 4.6. The covariance matrix of the roundoff error is no longer diagonal, since $\boldsymbol{C}^{(1)^{\mathrm{T}}}\boldsymbol{C}^{(1)} = 2\boldsymbol{I}_8 + 2\boldsymbol{J}_8 + 2\boldsymbol{K}_8$ with $\boldsymbol{K}_8 := \begin{pmatrix} \boldsymbol{J}_4 & \boldsymbol{I}_4 \\ \boldsymbol{I}_4 & \boldsymbol{J}_4 \end{pmatrix}$, $\boldsymbol{C}^{(2)^{\mathrm{T}}}\boldsymbol{C}^{(2)} = \boldsymbol{C}^{(3)^{\mathrm{T}}}\boldsymbol{C}^{(3)} = \boldsymbol{I}_8 + \boldsymbol{J}_8 - \boldsymbol{K}_8$ and $\boldsymbol{C}^{(j)^{\mathrm{T}}}\boldsymbol{C}^{(j)} = \boldsymbol{I}_8 - \boldsymbol{J}_8$ for $j = 4, 5, 6, 7$. By similar arguments we obtain (for cascade summation) that

$$\mathrm{Cov}\,(\boldsymbol{\Delta}) = \Big(3\sigma_{\mathbb{R},+}^2 + \sigma_{\mathbb{R},\times}^2 + \frac{101}{768}\Big)\rho^2 u^2 \boldsymbol{I}_8 + \frac{1}{256}\rho^2 u^2 \boldsymbol{J}_8 + \frac{1}{64}\rho^2 u^2 \boldsymbol{K}_8\,,$$

where the fractional constants follow from the identities $\mathbb{E}(2|\delta_1^C|^2 + \sum_{j=2}^{7}|\delta_j^C|^2) = \frac{101}{768}u^2$, $\mathbb{E}(2|\delta_1^C|^2 + |\delta_2^C|^2 + |\delta_3^C|^2 - \sum_{j=4}^{7}|\delta_j^C|^2) = \frac{1}{256}u^2$ and $\mathbb{E}(2|\delta_1^C|^2 - |\delta_2^C|^2 - |\delta_3^C|^2) = \frac{1}{64}u^2$. The norm is the same as for the naive DCT-II in Example 4.6.

4.9 Example. The fast DCT-III for $n = 8$ proposed by Loeffler et al. [9] is based on the factorization $\boldsymbol{C}_8^{\mathrm{III}} = 8^{-1/2}\boldsymbol{L}^{(1)}\boldsymbol{L}^{(2)^{\mathrm{T}}}\boldsymbol{L}^{(3)}\boldsymbol{L}^{(4)^{\mathrm{T}}}\boldsymbol{P}^{\mathrm{T}}$ with the notations from Example 4.7. The covariance matrices $\boldsymbol{K}^{(j)}$ $(j = 1, 2, 3, 4)$ of the roundoff error vectors $\boldsymbol{\Delta}^{(j)}$ resulting from the multiplication by $\boldsymbol{L}^{(j)^{\mathrm{T}}}$ remain essentially the same; only the leading factors $2\rho^2, 2\rho^2, 4\rho^2, 4\rho^2$ in Example 4.7 have to be replaced by $8\rho^2, 4\rho^2, 4\rho^2, \rho^2$ due to the different variance of the intermediate results. A simple calculation yields

$$\boldsymbol{L}^{(1)}\boldsymbol{K}^{(2)}\boldsymbol{L}^{(1)^{\mathrm{T}}} = \rho^2\Big[2(k_1+k_3)\boldsymbol{I}_8 - 2(k_1+k_3)\boldsymbol{J}_8 + 2(k_3-k_1)\boldsymbol{N}_8\Big]u^2,$$

$$\begin{aligned}\boldsymbol{L}^{(1)}\boldsymbol{L}^{(2)^{\mathrm{T}}}\boldsymbol{L}^{(3)}\boldsymbol{K}^{(4)}\boldsymbol{L}^{(3)}\boldsymbol{L}^{(2)}\boldsymbol{L}^{(1)} = \rho^2\Big[&2\sigma_{\mathbb{R},+}^2\Big(2\boldsymbol{I}_8 + \boldsymbol{R}_8 + \frac{\sqrt{2}}{2}\boldsymbol{Q}_8\Big)\\ &+ k_0(\boldsymbol{I}_8 + \boldsymbol{J}_8 - \boldsymbol{R}_8) + k_2\Big(\boldsymbol{I}_8 - \boldsymbol{J}_8 - \frac{\sqrt{2}}{2}\boldsymbol{Q}_8\Big)\Big]u^2,\end{aligned}$$

where $\boldsymbol{N}_8 := \mathrm{diag}(1,-1,-1,1,1,-1,-1,1)(\boldsymbol{I}_8 - \boldsymbol{J}_8)$, $\boldsymbol{R}_8 := \begin{pmatrix} \boldsymbol{J}_4 & \boldsymbol{I}_4 \\ \boldsymbol{I}_4 & \boldsymbol{J}_4 \end{pmatrix}$, and

$$\boldsymbol{Q}_8 := \begin{pmatrix} 0 & 1 & 1 & 0 & 0 & -1 & -1 & 0 \\ 1 & 0 & 0 & 1 & -1 & 0 & 0 & -1 \\ 1 & 0 & 0 & -1 & 1 & 0 & 0 & -1 \\ 0 & 1 & -1 & 0 & 0 & 1 & -1 & 0 \\ 0 & -1 & 1 & 0 & 0 & -1 & 1 & 0 \\ -1 & 0 & 0 & 1 & -1 & 0 & 0 & 1 \\ -1 & 0 & 0 & -1 & 1 & 0 & 0 & 1 \\ 0 & -1 & -1 & 0 & 0 & 1 & 1 & 0 \end{pmatrix}.$$

Therefore the total error

$$\boldsymbol{\Delta} := 8^{-1/2}\Big(\boldsymbol{\Delta}^{(1)} + \boldsymbol{L}^{(1)}\boldsymbol{\Delta}^{(2)} + \boldsymbol{L}^{(1)}\boldsymbol{L}^{(2)^{\mathrm{T}}}\boldsymbol{\Delta}^{(3)} + \boldsymbol{L}^{(1)}\boldsymbol{L}^{(2)^{\mathrm{T}}}\boldsymbol{L}^{(3)}\boldsymbol{\Delta}^{(4)}\Big)$$

has the covariance matrix

$$\begin{aligned}\mathrm{Cov}\,(\boldsymbol{\Delta}) = \Big[&\Big(\frac{13}{4}\sigma_{\mathbb{R},+}^2 + \sigma_{\mathbb{R},\times}^2 + \frac{121}{768}\Big)\boldsymbol{I}_8 - \Big(\frac{1}{4}\sigma_{\mathbb{R},+}^2 + \frac{1}{2}\sigma_{\mathbb{R},\times}^2 + \frac{41}{768}\Big)\boldsymbol{J}_8)\\ &+ \frac{5}{256}\,\boldsymbol{N}_8 - \frac{\sqrt{2}}{8}\Big(\sigma_{\mathbb{R},\times}^2 - \sigma_{\mathbb{R},+}^2 + \frac{1}{6}\Big)\boldsymbol{Q}_8 - \Big(\frac{1}{4}\sigma_{\mathbb{R},\times}^2 + \frac{5}{96}\Big)\boldsymbol{R}_8\Big]\rho^2 u^2,\end{aligned}$$

which leads to the same norm estimate as for the fast DCT-II algorithm. Again, the fractional constants above are the estimates for the precomputation errors. If these are all known exactly, the precision of our result can be improved. For example the constant $\frac{121}{768}u^2$ occuring in the norm estimate can be replaced by the quantity $\frac{1}{8}\big(|\delta^{\sqrt{2}}|^2+|\delta^{\sqrt{2}\sin\frac{3\pi}{8}}|^2+|\delta^{\sqrt{2}\cos\frac{3\pi}{8}}|^2+2|\delta^{\sin\frac{\pi}{16}}|^2+2|\delta^{\cos\frac{\pi}{16}}|^2+2|\delta^{\sin\frac{3\pi}{16}}|^2+2|\delta^{\cos\frac{3\pi}{16}}|^2\big)$, if we know the exact sizes of these errors.

4.10 Remark. In order to check the accuracy of our results by a simulation, we perform a discrete Fourier (or cosine) transformation followed by its inverse and compare the result with the original data. If $\boldsymbol{K}$ is the covariance matrix of the roundoff vector of the first unitary transformation, $\boldsymbol{X} \mapsto \boldsymbol{AX}$ and $\boldsymbol{K}'$ that of the second, $\boldsymbol{X} \mapsto A^*\boldsymbol{X}$ and if the roundoff errors of the two transformations are independent, then the observed roundoff error $\mathrm{fl}(A^*\,\mathrm{fl}(\boldsymbol{AX}))-\boldsymbol{X}$ has the covariance matrix $\tilde{\boldsymbol{K}} = \boldsymbol{K}' + \boldsymbol{A}^*\boldsymbol{K}\boldsymbol{A}$.

For the fast DCT-II (see Example 4.7) followed by the fast DCT-III (see Example 4.9), we have $\tilde{\boldsymbol{K}} = 2\boldsymbol{K}'$. In the following diagrams of Figure 4 we compare this estimation with a simulation with 100000 samples in MATLAB. We have not used the built–in matrix–vector multiplication, which in MATLAB is implemented differently from the standard algorithm.

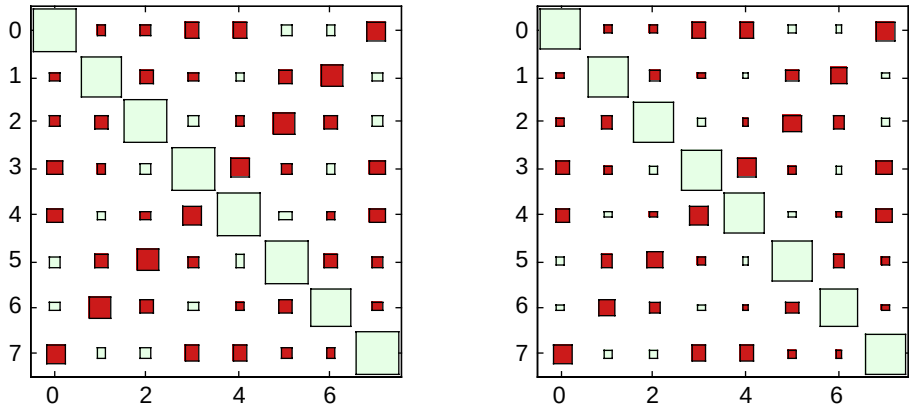


Figure 4: Computed (right) and simulated (left) covariance matrix for the roundoff errors of the fast DCT-II, followed by the fast DCT-III. The areas are proportional to the absolute values of the covariance and negative covariances are depicted by a darker shading.

The minor difference between the two covariances (particularly outside the diagonal) stem from two effects: firstly, since the precomputed constants are the same for every simulation, we observe the covariance conditioned on these values. Secondly the same constants appear in both the DCT-II and the DCT-III. This depencency increases their contribution to the covariance matrix $\tilde{\boldsymbol{K}}$ by up to a factor 2.

References

[1] M. Arioli, H. Munthe-Kaas and L. Valdettaro, *Componentwise error analysis for* FFT*s with applications to fast Helmholtz solvers*, Numer. Algorithms **12** (1996), 65–88.

[2] G. Baszenski, U. Schreiber, M. Tasche, *Numerical stability of fast cosine transforms*, Numer. Funct. Anal. Optim. **21** (2000), 25–46.

[3] J. Bustoz, A. Feldstein, R. Goodman, S. Linnainmaa, *Improved trailing digits estimates applied to optimal computer arithmetic*, J. ACM **26** (1979), 716–730.

[4] D. Calvetti, *A stochastic roundoff error analysis for the fast Fourier transform*, Math. Comp. **56** (1991), 755 – 774.

[5] W. Chen, C. Harrison, and S. Fralick, *A fast computational algorithm for the discrete cosine transform*, IEEE Trans. Comput. **25** (1977), 1004 – 1011.
[6] I. Daubechies, W. Sweldens, *Factoring wavelet transforms into lifting steps*, J. Fourier Anal. Appl. **4** (1998), 247 – 269.
[7] N. J. Higham, *Accuracy and Stability of Numerical Algorithms*, SIAM, Philadelphia, 1996.
[8] C. Van Loan, *Computational Frameworks for the Fast Fourier Transform*, SIAM, Philadelphia, 1992.
[9] C. Loeffler, A. Lightenberg, and G. Moschytz, *Practical fast 1-d DCT algorithms with 11 multiplications*, Proc. IEEE ICASSP **2** (1989), 988 – 991.
[10] H. J. Nussbaumer, *Fast Fourier Transform and Convolution Algorithms*, Springer–Verlag, Berlin, 1982.
[11] J. C. Schatzman, *Accuracy of the discrete Fourier transform and the fast Fourier transform*, SIAM J. Sci. Comput. **17** (1996), 1150 – 1166.
[12] U. Schreiber, *Fast and numerically stable trigonometric transforms*, PhD thesis, University of Rostock (1999), in German.
[13] M. Tasche, Hm. Zeuner, *Roundoff error analysis for the fast Fourier transform with precomputed twiddle factors*, Preprint, Medical University of Luebeck (1999).
[14] M. Tasche, Hm. Zeuner, *Improved roundoff error analysis for precomputed twiddle factors*, J. Comput. Anal. Appl. **4** (2002), in print.
[15] M. Tasche, Hm. Zeuner, *Worst and average case roundoff error analysis for FFT*, BIT **41** (2001), 563–581.

JOURNAL OF CONCRETE AND APPLICABLE MATHEMATICS,VOL.3,NO.3,313-334,2005,COPYRIGHT 2005 EUDOXUS PRESS,LLC

Differential Properties of Matrix Orthogonal Polynomials

M.J. Cantero, L. Moral, L.Velázquez

Departamento de Matemática Aplicada. Universidad de Zaragoza. Spain.

mjcante@unizar.es, lmoral@unizar.es, velazque@unizar.es

Abstract

In this paper a general theory of semi-classical matrix orthogonal polynomials is developed. We define the semi-classical linear functionals by means of a distributional equation $D(uA) = uB$, where A and B are matrix polynomials. Several characterizations for these semi-classical functionals are given in terms of the corresponding (left) orthogonal matrix polynomial sequence. They involve a quasi-orthogonality property for their derivatives, a structure relation and a second order differo-differential equation. Finally we illustrate the preceding results with some non-trivial examples.

Keywords and phrases: Orthogonal matrix polynomials, Semi-classical functionals, Quasi-orthogonality, Structure relation, Differential equation.

(2000) AMS Mathematics Subject Classification: 42C05

1 Introduction

The study of semi-classical orthogonal polynomials in the scalar case (i.e., the study of orthogonal polynomials whose associated functional satisfies a distributional equation $D(u\Phi) = u\Psi$, where Φ, Ψ are polynomials with $\deg \Psi \geq 1$) was started by Shohat ([14]) in order to generalize the properties of the classical orthogonal polynomials ([3]).

Among others, [2, 9] present an approach to such polynomials taking into account the quasi-orthogonality of the derivatives of the polynomial sequence. We also mention the results of Maroni who, in [11, 12], develops an algebraic theory of semi-classical orthogonal polynomials. The purpose of this theory is the characterization of the semi-classical orthogonal polynomials by means of the quasi-orthogonality of their derivatives, a structure relation and a differo-differential equation of second order (see also [1]).

In the last years, the study of matrix orthogonal polynomials has attracted the interest of the researchers (see [4, 6, 8, 10]). As for their differential properties, it is known the result of Durán ([5]) who characterizes those matrix orthogonal polynomials satisfying a symmetric second order differential equation with polynomial coefficients. He proves that, in the positive definite case, if the differential equation has its coefficients on the same side, the matrix orthogonal polynomials are necessarily diagonal (up to factors) with classical scalar orthogonal polynomials in the diagonal. In spite of the variety of applications of matrix polynomials ([4, 5, 6]), there are not too many known families of semi-classical matrix orthogonal polynomials out of the diagonal case.

One way to study many families of matrix orthogonal polynomials is to extend the analysis of differential properties started by Durán. A natural way to do this is to generalize the theory of semi-classical scalar orthogonal polynomials to the matricial case. In order to do that, we start defining the semi-classical matrix orthogonal polynomials by a Pearson-type equation for the corresponding orthogonality matrix functional. We prove that such matrix orthogonal polynomials are characterized by a (in general, non-symmetric) differo-differential equation. In the way to the proof, we find another equivalences that generalize the known ones in the scalar case.

This paper has been organized as follows. In Section 3 we introduce and study the concept of quasi-orthogonality for matrix polynomials.

The main result of Section 4 is the characterization of the semi-classical functionals in terms of the quasi-orthogonality of the corresponding matrix orthogonal polynomials and their derivatives.

In Section 5 we obtain for the semi-classical matrix orthogonal polynomials a structure relation and a differo-differential equation, showing that they are equivalent to a Pearson-type equation for the related orthogonality matrix functional.

Finally, in Section 6, we illustrate the preceding results with some examples, showing that there exist non-diagonalizable semi-classical matrix functionals.

2 Basic tools

Denoting by $\mathbb{C}^{(m,m)}$ the set of $m \times m$ complex matrices, in what follows $\mathbb{P}^{(m)}$ will be the $\mathbb{C}^{(m,m)}$-left-module

$$\mathbb{P}^{(m)} = \left\{ \sum_{k=0}^{n} \alpha_k x^k \,\middle|\, \alpha_k \in \mathbb{C}^{(m,m)},\ n \in \mathbb{N} \right\}$$

and $\mathbb{P}^{(m)'}$ the $\mathbb{C}^{(m,m)}$-right-module $\mathrm{Hom}\big(\mathbb{P}^{(m)}, \mathbb{C}^{(m,m)}\big)$.

2.1 Definition

(i) The duality bracket is defined by

$$\begin{aligned}\big\langle \cdot\ ,\ \cdot \big\rangle :\ \mathbb{P}^{(m)} \times \mathbb{P}^{(m)'} &\to \mathbb{C}^{(m,m)} \\ \big(P, u\big) &\to \big\langle P, u\big\rangle := u(P)\end{aligned}$$

(ii) Given $u \in \mathbb{P}^{(m)'}$, the functional $ux^k I \in \mathbb{P}^{(m)'}$ $(k = 0, 1, \ldots)$ is defined by

$$\big\langle P, ux^k I\big\rangle := \big\langle x^k P, u\big\rangle,$$

where I denotes the $m \times m$ identity matrix. A linear extension gives the right-product $uA \in \mathbb{P}^{(m)'}$ for any $A \in \mathbb{P}^{(m)}$, with $A(x) = \sum_{k=0}^{n} \alpha_k x^k$, in the following way:

$$\big\langle P, uA\big\rangle = \sum_{k=0}^{n} \big\langle x^k P, u\big\rangle \alpha_k.$$

(iii) Every $u \in \mathbb{P}^{(m)'}$, defines a matrix inner product in $\mathbb{P}^{(m)}$ by

$$\begin{aligned}\big\langle \cdot\ ,\ \cdot \big\rangle_u :\ \mathbb{P}^{(m)} \times \mathbb{P}^{(m)} &\to \mathbb{C}^{(m,m)} \\ \big(P, Q\big) &\to \big\langle P, Q\big\rangle_u := \big\langle P, uQ^*\big\rangle\end{aligned}$$

where Q^* denotes the transposed conjugated of Q.

Remark The duality bracket verifies the property of linearity

$$\begin{aligned}&\big\langle \alpha_1 P_1 + \alpha_2 P_2, \beta_1 u_1 + \beta_2 u_2\big\rangle = \\ &\qquad = \alpha_1 \big\langle P_1, u_1\big\rangle \beta_1 + \alpha_1 \big\langle P_1, u_2\big\rangle \beta_2 + \alpha_2 \big\langle P_2, u_1\big\rangle \beta_1 + \alpha_2 \big\langle P_2, u_2\big\rangle \beta_2,\end{aligned}$$

for all $\alpha_1, \alpha_2, \beta_1, \beta_2 \in \mathbb{C}^{(m,m)}$, $P_1, P_2 \in \mathbb{P}^{(m)}$ and $u_1, u_2 \in \mathbb{P}^{(m)'}$, while the inner product is sesquilinear, that is,

$$\begin{aligned}&\big\langle \alpha_1 P_1 + \alpha_2 P_2, \beta_1 Q_1 + \beta_2 Q_2\big\rangle_u = \\ &\qquad = \alpha_1 \big\langle P_1, Q_1\big\rangle_u \beta_1^* + \alpha_1 \big\langle P_1, Q_2\big\rangle_u \beta_2^* + \alpha_2 \big\langle P_2, Q_1\big\rangle_u \beta_1^* + \alpha_2 \big\langle P_2, Q_2\big\rangle_u \beta_2^*,\end{aligned}$$

for all $\alpha_1, \alpha_2, \beta_1, \beta_2 \in \mathbb{C}^{(m,m)}$ and $P_1, P_2, Q_1, Q_2 \in \mathbb{P}^{(m)}$.

2.2 Definition *We denote by $C_k := \big\langle x^k I, u\big\rangle$ the k-th moment with respect to $u \in \mathbb{P}^{(m)'}$. Given $u \in \mathbb{P}^{(m)'}$ with moments $(C_k)_{k=0}^{\infty}$, we say that u is quasi-definite (non-singular) if $\det\big(C_{k+j}\big)_{k,j=0}^{n} \neq 0$, $\forall n \geq 0$, where $(C_{k+j})_{k,j=0}^{n}$ is the Hankel-block matrix*

$$\begin{pmatrix} C_0 & C_1 & \ldots & C_n \\ C_1 & C_2 & \ldots & C_{n+1} \\ \ldots & \ldots & \ldots & \ldots \\ C_n & C_{n+1} & \ldots & C_{2n} \end{pmatrix}.$$

Remark Given a sequence $\left(C_k\right)_{k=0}^{\infty} \subset \mathbb{C}^{(m,m)}$, there exists a unique $u \in \mathbb{P}^{(m)'}$ such that $\left\langle x^k I, u\right\rangle = C_k$. This establishes an isomorphism between $\mathbb{P}^{(m)'}$ and the formal series $\sum_{k=0}^{\infty} C_k x^k$ with coefficients in $\mathbb{C}^{(m,m)}$.

2.3 Definition *We say that $u \in \mathbb{P}^{(m)'}$ is hermitian if $C_k^* = C_k$ for all $k \geq 0$.*

The following result is very well known (see [4, 6, 7, 8, 13]).

2.4 Theorem *$u \in \mathbb{P}^{(m)'}$ is quasi-definite if and only if there exists a unique (up to non-singular left matrix factors) sequence of left orthogonal matrix polynomials $\left(P_n\right)_{n\geq 0}$ with respect to u, that is:*

(1) $P_n \in \mathbb{P}^{(m)}$, $\deg P_n = n$.

(2) The leading coefficient of P_n is non-singular.

(3) $\left\langle P_n, P_k\right\rangle_u = K_n \delta_{nk}$ *for $k \leq n$, where K_n is non-singular.*

This sequence verifies a (left) recurrence relation

$$\begin{aligned} &xP_n(x) = \alpha_n P_{n+1}(x) + \beta_n P_n(x) + \gamma_n P_{n-1}(x), \qquad n \geq 0, \\ &P_{-1}(x) = \theta, \end{aligned} \tag{1.1}$$

where α_n, β_n, $\gamma_n \in \mathbb{C}^{(m,m)}$, with α_n, γ_n non-singular, and θ denotes the zero $m \times m$ matrix.

Remarks

(i) The inner product $\left\langle P_n, x^n I\right\rangle_u = \left\langle x^n P_n, u\right\rangle$ is non-singular for every quasi-definite functional $u \in \mathbb{P}^{(m)'}$.

(ii) If we consider the structure of right-module on $\mathbb{P}^{(m)}$ (so, left-module on $\mathbb{P}^{(m)'}$), we can define in a similar way a sequence $(\Pi_n)_{n\geq 0}$ of right orthogonal matrix polynomials with respect to a quasi-definite matrix functional u. The difference is that the condition (3) of Theorem 2.4 must be substituted by $\left\langle \Pi_k^*, \Pi_n^*\right\rangle_u = L_n \delta_{nk}$ for $k \leq n$, with L_n non-singular. The sequence $(\Pi_n)_{n\geq 0}$ satisfies a right recurrence relation.

3 Quasi-orthogonality

As in the scalar case, the semi-classical character will be closely related to the idea of quasi-orthogonality. But the scalar quasi-orthogonality, when generalized to the matricial case, splits in two different concepts. Both of them will play an important role in the characterization of the semi-classical matrix functionals.

3.1 Definition *Let $\left(Q_n\right)_{n\geq 0}$ be a sequence of matrix polynomials such that the leading coefficient of Q_n is non-singular and* $\deg Q_n = n$.

1. *We will say that $\left(Q_n\right)_{n\geq 0}$ is a sequence of left quasi-orthogonal matrix polynomials of order r with respect to $v \in \mathbb{P}^{(m)'}$ if*

(a) $\left\langle x^k Q_n, v\right\rangle = \theta, \quad 0 \leq k \leq n-r-1, \quad n \geq r+1,$

(b1) There exists $n_0 \geq r$ *such that* $\left\langle x^{n_0-r} Q_{n_0}, v\right\rangle \neq \theta$.

2. *We will say that $\left(Q_n\right)_{n\geq 0}$ is a sequence of regularly left quasi-orthogonal matrix polynomials of order r with respect to $v \in \mathbb{P}^{(m)'}$ if*

(a) $\left\langle x^k Q_n, v\right\rangle = \theta, \quad 0 \leq k \leq n-r-1, \quad n \geq r+1,$

(b2) There exists $n_0 \geq r$ *such that* $\left\langle x^{n_0-r} Q_{n_0}, v\right\rangle$ *is non-singular.*

Notice that if $r = 0$ the previous definition is the standard orthogonality. Furthermore, for $m = 1$ both definitions yield the scalar quasi-orthogonality.

As we will see in the next proposition, when a sequence of quasi-orthogonal matrix polynomials is already orthogonal with respect to another functional, the conditions (b1), (b2) become stronger.

3.2 Proposition *Let $u \in \mathbb{P}^{(m)'}$ be quasi-definite and let $\left(P_n\right)_{n\geq 0}$ be the corresponding sequence of left orthogonal matrix polynomials. Given $v \in \mathbb{P}^{(m)'}$, the sequence $\left(P_n\right)_{n\geq 0}$ is quasi-orthogonal of order r with respect to v if and only if*

(a) $\left\langle x^k P_n, v\right\rangle = \theta, \quad 0 \leq k \leq n-r-1, \quad n \geq r+1,$

(b1') $\left\langle x^{n-r} P_n, v\right\rangle \neq \theta, \quad \forall n \geq r.$

The sequence $\left(P_n\right)_{n\geq 0}$ is regularly quasi-orthogonal of order r with respect to v if and only if

(a) $\left\langle x^k P_n, v\right\rangle = \theta, \quad 0 \leq k \leq n-r-1, \quad n \geq r+1,$

(b2') $\langle x^{n-r}P_n, v\rangle$ *is non-singular* $\forall n \geq r$.

Proof Let $n \geq r$ verifying (b1) or (b2). Taking into account the recurrence relation (1.1), we obtain

$$\gamma_{n+1}\langle x^{n-r}P_n, v\rangle = \langle x^{n+1-r}P_{n+1}, v\rangle$$

and (b1) or (b2) holds for $n+1$. In a similar way, if $n \geq r+1$,

$$\gamma_n\langle x^{n-1-r}P_{n-1}, v\rangle = \langle x^{n-r}P_n, v\rangle$$

and (b1) or (b2) is satisfied for $n-1$. □

Remember that, given $w \in \mathbb{P}^{(m)'}$, we say that $w = 0$ if $\langle P, w\rangle = \theta$, $\forall P \in \mathbb{P}^{(m)}$. In what follows, the following result will be useful.

3.3 Lemma *Let $u \in \mathbb{P}^{(m)'}$ be quasi-definite and let $(P_n)_{n\geq 0}$ be the corresponding sequence of left orthogonal matrix polynomials. Then, given $w \in \mathbb{P}^{(m)'}$, $w = 0$ if and only if there exists n_0 such that*

$$\langle x^k P_n, w\rangle = \theta, \qquad 0 \leq k \leq n, \qquad n \geq n_0.$$

Proof Assume the above orthogonality conditions. From the recurrence (1.1),

$$\langle x^k P_{n_0-1}, w\rangle = \gamma_{n_0}^{-1}\left(\langle x^{k+1}P_{n_0}, w\rangle - \alpha_{n_0}\langle x^k P_{n_0+1}, w\rangle - \beta_{n_0}\langle x^k P_{n_0}, w\rangle\right) = \theta$$

for $0 \leq k \leq n_0 - 1$. So, the hypothesis is true for $n \geq n_0 - 1$, $0 \leq k \leq n$, too, and, by induction, for $0 \leq k \leq n$. Therefore, $\langle P_n, w\rangle = \theta$ for all $n \geq 0$ and, thus, $w = 0$. □

Given a sequence of orthogonal matrix polynomials with respect to a functional u, its quasi-orthogonality with respect to another functional v can be characterized by a simple relation between u and v.

3.4 Theorem *Let $u, v \in \mathbb{P}^{(m)'}$ such that u is quasi-definite and let $(P_n)_{n\geq 0}$ be the corresponding sequence of left orthogonal matrix polynomials. Then,*

(i) *$(P_n)_{n\geq 0}$ is quasi-orthogonal of order r with respect to v if and only if there exists $A \in \mathbb{P}^{(m)}$, with $\deg A = r$, such that $v = uA$.*

(ii) *$(P_n)_{n\geq 0}$ is regularly quasi-orthogonal of order r with respect to v if and only if there exists $A \in \mathbb{P}^{(m)}$, with $\deg A = r$ and non-singular leading coefficient, such that $v = uA$.*

Moreover, in any case, the matrix polynomial A is unique.

Proof We will do the proof just for (i) because for (ii) the arguments are analogous.

($\Leftarrow$) Consider $A(x)=\sum_{j=0}^{r} a_j x^j$, with $a_r \neq \theta$. Then, if $v = uA$,

$$\left\langle x^k P_n, v\right\rangle = \sum_{j=0}^{r} \left\langle x^{k+j} P_n, u\right\rangle a_j.$$

Hence, for $0 \leq k \leq n-r-1$, we have

$$\left\langle x^{k+j} P_n, u\right\rangle = \theta, \qquad 0 \leq j \leq r,$$

and, for $k = n-r$,

$$\left\langle x^k P_n, v\right\rangle = \langle x^n P_n, u\rangle\, a_r \neq \theta.$$

So, $(P_n)_{n\geq 0}$ is quasi-orthogonal of order r with respect to v.

($\Rightarrow$) We will prove that $0 = v - \sum_{j=0}^{r} x^j u a_j$ has a unique solution $(a_j)_{j=0}^{r}$.

For all $n \geq r$, the system of $r+1$ equations

$$\begin{cases} \left\langle x^{n-r} P_n, v\right\rangle = \langle x^n P_n, u\rangle\, a_r^{(n)}, \\ \qquad\qquad \vdots \\ \langle x^n P_n, v\rangle = \langle x^n P_n, u\rangle\, a_0^{(n)} + \cdots + \left\langle x^{n+r} P_n, u\right\rangle a_r^{(n)}, \end{cases} \tag{2.1}$$

has a unique solution in the unknowns $(a_j^{(n)})_{j=0}^{r}$. Notice that the first equation leads to $a_r^{(n)} \neq \theta$.

We will prove that $(a_j^{(n)})_{j=0}^{r}$ are independent of n. From (1.1), the relation

$$\gamma_{n+1}\left\langle x^k P_n, \hat{v}\right\rangle + \beta_{n+1}\left\langle x^k P_{n+1}, \hat{v}\right\rangle + \alpha_{n+1}\left\langle x^k P_{n+2}, \hat{v}\right\rangle = \left\langle x^{k+1} P_{n+1}, \hat{v}\right\rangle \tag{2.2}$$

is verified for all $k \geq 0$ and for all $\hat{v} \in \mathbb{P}^{(m)'}$. The first equation of (2.1) gives

$$\begin{cases} \left\langle x^{n-r} P_n, v\right\rangle = \langle x^n P_n, u\rangle\, a_r^{(n)} \\ \left\langle x^{n+1-r} P_{n+1}, v\right\rangle = \left\langle x^{n+1} P_{n+1}, u\right\rangle a_r^{(n+1)}, \end{cases} \tag{2.3}$$

and, from (2.2),

$$\begin{cases} \gamma_{n+1}\left\langle x^{n-r} P_n, v\right\rangle - \left\langle x^{n+1-r} P_{n+1}, v\right\rangle = \theta, \\ \gamma_{n+1}\left\langle x^n P_n, u\right\rangle - \left\langle x^{n+1} P_{n+1}, u\right\rangle = \theta. \end{cases} \tag{2.4}$$

Equations (2.3) and (2.4) imply that $a_r^{(n)} = a_r^{(n+1)}$.

Suppose $a_j^{(n)} = a_j^{(n+1)}$ for $j = k+1, \dots, r$, with $k < r$, and for all $n \geq r$. We will prove that $a_k^{(n)} = a_k^{(n+1)}$ too. Taking into account (2.1) we have

$$\begin{cases} \left\langle x^{n-k} P_n, v \right\rangle = \sum_{j=k}^{r} \left\langle x^{n-k+j} P_n, u \right\rangle a_j^{(n)}, \\ \left\langle x^{n-k+1} P_{n+1}, v \right\rangle = \sum_{j=k}^{r} \left\langle x^{n+1-k+j} P_{n+1}, u \right\rangle a_j^{(n+1)}, \end{cases} \tag{2.5}$$

and, from (2.2),

$$\gamma_{n+1}\left\langle x^{n-k} P_n, v \right\rangle + \beta_{n+1}\left\langle x^{n-k} P_{n+1}, v \right\rangle + \alpha_{n+1}\left\langle x^{n-k} P_{n+2}, v \right\rangle = \\ = \left\langle x^{n-k+1} P_{n+1}, v \right\rangle, \tag{2.6}$$

$$\gamma_{n+1}\left\langle x^{n-k+j} P_n, u \right\rangle + \beta_{n+1}\left\langle x^{n-k+j} P_{n+1}, u \right\rangle + \alpha_{n+1}\left\langle x^{n-k+j} P_{n+2}, u \right\rangle = \\ = \left\langle x^{n+1-k+j} P_{n+1}, u \right\rangle. \tag{2.7}$$

Substituting (2.5) into (2.6) we obtain

$$\gamma_{n+1} \sum_{j=k}^{r} \left\langle x^{n-k+j} P_n, u \right\rangle a_j^{(n)} + \beta_{n+1} \sum_{j=k+1}^{r} \left\langle x^{n-k+j} P_{n+1}, u \right\rangle a_j^{(n+1)} + \\ + \alpha_{n+1} \sum_{j=k+2}^{r} \left\langle x^{n-k+j} P_{n+2}, u \right\rangle a_j^{(n+2)} = \sum_{j=k}^{r} \left\langle x^{n+1-k+j} P_{n+1}, u \right\rangle a_j^{(n+1)}.$$

From this equality, taking into account (2.7) and applying the induction hypothesis, we get

$$\gamma_{n+1} \left\langle x^n P_n, u \right\rangle a_k^{(n)} - \left\langle x^{n+1} P_{n+1}, u \right\rangle a_k^{(n+1)} = \theta.$$

Thus, by (2.4), we conclude that $a_k^{(n)} = a_k^{(n+1)}$.

Let us denote by $(a_j)_{j=0}^r$ the solutions of (2.1) for $n \geq r$, which we have proved that are independent of n. If $A(x) = \sum_{j=0}^{r} a_j x^j$ and $w = v - uA$,

$$\left\langle x^k P_n, w \right\rangle = \left\langle x^k P_n, v \right\rangle - \sum_{j=0}^{r} \left\langle x^{k+j} P_n, u \right\rangle a_j, \quad 0 \leq k \leq n, \quad n \geq r.$$

When $0 \leq k \leq n-r-1$, the orthogonality with respect to u and the quasi-ortogonality with respect to v give $\left\langle x^k P_n, w \right\rangle = \theta$. Furthermore, $\left\langle x^k P_n, w \right\rangle = \theta$

for $n-r \leq k \leq n$ too, because $(a_j)_{j=0}^r$ are the solutions of the equations (2.1). So, $w=0$, according to Proposition 2.2.

To prove the uniqueness of the matrix polynomial A, let us suppose $A, B \in \mathbb{P}^{(m)}$ such that $v = uA = uB$. Then, from (i), $r = \deg A = \deg B$ since $(P_n)_{n\geq 0}$ can not be quasi-orthogonal of different orders with respect to v. Hence, $A(x) = \sum_{j=0}^{r} a_j x^j$, $B(x) = \sum_{j=0}^{r} b_j x^j$ with $a_r, b_r \neq \theta$. From

$$\left\langle x^k P_n, v\right\rangle = \sum_{j=0}^{r} \left\langle x^{k+j} P_n, u\right\rangle a_j = \sum_{j=0}^{r} \left\langle x^{k+j} P_n, u\right\rangle b_j,$$

and taking $k = n-r, \ldots, n$, we find

$$\begin{cases} \langle x^n P_n, u\rangle (a_r - b_r) = \theta, \\ \qquad\qquad \vdots \\ \left\langle x^{n+r} P_n, u\right\rangle (a_r - b_r) + \cdots + \langle x^n P_n, u\rangle (a_0 - b_0) = \theta. \end{cases}$$

Therefore, $a_j = b_j$ for $j = 0, \cdots, r$. □

4 Semi-classical functionals

We consider the derivative operator on the space $\mathbb{P}^{(m)'}$ as the linear operator $D : \mathbb{P}^{(m)'} \to \mathbb{P}^{(m)'}$ such that $\langle P, Du\rangle = -\langle P', u\rangle$. From this and the definition of the right-product it is straightforward to prove that $D(uA) = (Du)A + uA'$, for all $u \in \mathbb{P}^{(m)'}$ and $A \in \mathbb{P}^{(m)}$.

4.1 Definition *Let $u \in \mathbb{P}^{(m)'}$ be quasi-definite. We will say that u is semi-classical if there exist $A, B \in \mathbb{P}^{(m)}$, with $\det A \neq 0$, such that it is verified the distributional equation $D(uA) = uB$. We will also say that the corresponding sequence of left orthogonal matrix polynomials $(P_n)_{n\geq 0}$ is semi-classical.*

Remarks

(i) Given $u \in \mathbb{P}^{(m)'}$, the linear functionals $u_{ij} : \mathbb{P}^{(m)} \to \mathbb{C}$ $(i, j = 1, \cdots, m)$ defined by $u_{ij}(P) = u(P)_{ij}$, $\forall P \in \mathbb{P}^{(m)}$, are called the components of u. Then, $(uA)_{ij} = \sum_{k=1}^{m} u_{ik} A_{kj}$ with the obvious definition for the multiplication of u_{ij} by a scalar polynomial. Therefore, $\det A \neq 0$ is the minimal requirement

to ensure that the previous definition will involve to all the components of the functional u.

(ii) Notice that the matrix polynomial B of the Pearson-type equation of a semi-classical matrix functional can not be null. Otherwise, the Pearson-type equation implies $D(uA) = 0$ and, then, $\det(C_{k+j})_{k,j=0}^{p} = 0$, where $p = \deg A$. This contradicts the non-singularity of u.

4.2 Lemma *Let $u \in \mathbb{P}^{(m)'}$ such that $D(uA) = uB$. Then, for every $C \in \mathbb{P}^{(m)}$,*

$$D(uAC) = u(AC' + BC)$$

Proof It is just a consequence of the rule for the derivation of the right-product. □

The previous result implies that, for every $u \in \mathbb{P}^{(m)'}$, the set

$$\mathcal{A}_u = \{A \in \mathbb{P}^{(m)} \mid D(uA) = uB, B \in \mathbb{P}^{(m)}\}$$

is a right-ideal of $\mathbb{P}^{(m)}$.

When dealing with scalar semi-classical functionals we arrive in this way to an ideal of $\mathbb{P} := \mathbb{P}^{(1)}$, which is necessarily generated by a unique (up to non-trivial factors) polynomial. This generator is used to classify the semi-classical scalar functionals (see [11, 12]).

Since in the matricial case, a right-ideal of $\mathbb{P}^{(m)}$ is not necessarily principal, we cannot use $\mathcal{A}_u$ for the classification of semi-classical functionals. However, if $u \in \mathbb{P}^{(m)'}$ is semi-classical, then there exists $\alpha \in \mathbb{P} \setminus \{0\}$ such that $\alpha I \in \mathcal{A}_u$. To see this just notice that if $A \in \mathcal{A}_u$ with $\det A \neq 0$, then $(\det A)I \in \mathcal{A}_u$ since $(\det A)I = A \operatorname{adj} A$. Therefore, for every semi-classical matrix functional $u \in \mathbb{P}^{(m)'}$, the set

$$\vartheta_u = \{\alpha \in \mathbb{P} \mid D(u\alpha I) = uB, B \in \mathbb{P}^{(m)}\}$$

is a non-trivial ideal of $\mathbb{P}$. We can use the "essentially" unique generator of this ideal to classify the semi-classical matrix functionals similarly to the scalar case.

4.3 Definition *Let $u \in \mathbb{P}^{(m)'}$ be semi-classical and let $\alpha \in \mathbb{P}\setminus\{0\}$ be a polynomial with smallest degree such that $D(u\alpha I) = uB$, $B \in \mathbb{P}^{(m)}$. Then, we say that u is of class $s = \max\{p-2, q-1\}$, where $p = \deg \alpha$ and $q = \deg B$.*

Remarks

(i) The preceding discussion shows that this definition is well done: it always exists such a polynomial α and the definition of class does not depend on the choice of α.

(ii) When $u \in \mathbb{P}^{(m)'}$ is semi-classical there exists $A \in \mathcal{A}_u$ with $\det A \neq 0$, but it is possible for the leading coefficient of A to be singular. However, there always exists $\widetilde{A} \in \mathcal{A}_u$ with non-singular leading coefficient. To see this just take $\widetilde{A} = (\det A)I$.

The following theorem gives the first characterization of the semi-classical matrix functionals.

4.4 Theorem *Let $u \in \mathbb{P}^{(m)'}$ be quasi-definite and let $(P_n)_{n\geq 0}$ be the corresponding sequence of left orthogonal matrix polynomials. Then, the following statements are equivalent:*

(i) u is semi-classical.

(ii) There exists $v \in \mathbb{P}^{(m)'}$ such that $\left(P'_{n+1}\right)_{n\geq 0}$ is quasi-orthogonal with respect to v, and $(P_n)_{n\geq 0}$ is regularly quasi-orthogonal with respect to v.

Proof

$(i) \Rightarrow (ii)$ If u is semi-classical, then $D(u\alpha I) = uB$, $\alpha \in \mathbb{P} \setminus \{0\}$, $B \in \mathbb{P}^{(m)}$, according to the previous comments. Therefore,

$$\left\langle \left(x^k P_n\right)', u\alpha I \right\rangle + \left\langle x^k P_n, uB \right\rangle = \theta.$$

Let $\alpha(x) = \sum_{j=0}^{p} a_j x^j$, $B(x) = \sum_{j=0}^{q} b_j x^j$ with $a_p \neq 0$, $b_q \neq \theta$. With this notation the above equation becomes

$$\sum_{j=0}^{p} \left\langle k x^{k-1+j} P_n, u \right\rangle a_j + \sum_{j=0}^{q} \left\langle x^{k+j} P_n, u \right\rangle b_j = -\left\langle x^k P'_n, u\alpha I \right\rangle. \tag{3.1}$$

The left hand side of this expression is equal to the zero matrix if $0 \leq k \leq n-p$ and $0 \leq k \leq n-(q+1)$. Let $r = \max\{p-1, q\}$. Then, $\left\langle x^k P'_n, u\alpha I \right\rangle = 0$ if $0 \leq k \leq n-r-1$. For $k = n-r$, (3.1) becomes

$$(n-r)\left\langle x^n P_n, u \right\rangle a_{r+1} + \left\langle x^n P_n, u \right\rangle b_r = -\left\langle x^{n-r} P'_n, u\alpha I \right\rangle,$$

and thus

$$\left\langle x^{n-r} P'_n, u\alpha I \right\rangle = -\left\langle x^n P_n, u \right\rangle \left[(n-r)a_{r+1} + b_r\right],$$

which is equal to the zero matrix for at most one n. So, the sequence $(Q_n)_{n\geq 0}$, with $Q_n = P'_{n+1}$, is quasi-orthogonal of order $r-1$ with respect to $u\alpha I$.

Obviously, since αI has non-singular leading coefficient, $(P_n)_{n\geq 0}$ is regularly quasi-orthogonal of order p with respect to $u\alpha I$.

$(ii) \Rightarrow (i)$ If $(P_n)_{n\geq 0}$ is regularly quasi-orthogonal of order p with respect to v, there exists $A \in \mathbb{P}^{(m)}$, with $\deg A = p$ and non-singular leading coefficient, such that $v = uA$. Notice that if the leading coefficient of A is non-singular, then $\det A \neq 0$.

Let $w = D(uA)$. $\langle P, w\rangle = -\langle P', uA\rangle$ and for $P = x^k P_n$ we have

$$\left\langle x^k P_n, w\right\rangle = -k\left\langle x^{k-1}P_n, uA\right\rangle - \left\langle x^k P_n', uA\right\rangle.$$

From the quasi-orthogonality, the right hand side vanishes if $0 \leq k \leq n-p$ and $0 \leq k \leq n-s-2$, where s is the order of quasi-orthogonality of $\left(P_{n+1}'\right)_{n\geq 0}$. So, $(P_n)_{n\geq 0}$ is quasi-orthogonal with respect to the functional w with order at most $\max\{p-1, s+1\}$. Thus, there exists $B \in \mathbb{P}^{(m)}$ such that $\deg B \leq \max\{p-1, s+1\}$ and $w = uB$. □

Remark Notice that if $\left(P_{n+1}'\right)_{n\geq 0}$ is quasi-orthogonal of order s with respect to $u\alpha I$, then $D(u\alpha I) = uB$ where $\deg\alpha = p$, $\deg B = q$, and this implies that $s = \max\{p-2, q-1\}$.

5 Structure relation and differo-differential equation

5.1 Theorem Structure relation

Let $u \in \mathbb{P}^{(m)'}$ be quasi-definite and let $(P_n)_{n\geq 0}$ be the associated sequence of left orthogonal matrix polynomials. Then, the following statements are equivalent:

(i) u is semi-classical.

(ii) There exist a polynomial $\alpha \in \mathbb{P}\backslash\{0\}$ and matrices $\Theta_j^{(n)} \in \mathbb{C}^{(m,m)}$ $(n \geq 0, -s \leq j \leq p)$, such that

$$\alpha(x)P_{n+1}'(x) = \sum_{j=-s}^{p} \Theta_j^{(n)} P_{n+j}(x),$$

where $p = \deg\alpha$, $s \geq \max\{p-2, 0\}$ and $\Theta_{-s}^{(n)} \neq \theta$ for some $n \geq s$ (we use the convention $P_k = \theta$ for $k < 0$).

Proof

$(i) \Rightarrow (ii)$ If u is semi-classical, then there exist $\alpha \in \mathbb{P}\setminus\{0\}$ and $B \in \mathbb{P}^{(m)}$ such that $D(u\alpha I) = uB$. Hence, $\left(P_{n+1}'\right)_{n\geq 0}$ is quasi-orthogonal with respect to $u\alpha I$ of order $s = \max\{p-2, q-1\}$, where $p = \deg\alpha$ and $q = \deg B$. This implies

$$\left\langle x^k P_{n+1}', u\alpha I\right\rangle = \theta, \quad 0 \leq k \leq n-s-1,$$
$$\left\langle x^{n-s} P_{n+1}', u\alpha I\right\rangle \neq \theta, \quad \text{for some } n \geq s.$$

Since $\alpha P'_{n+1} \in \mathbb{P}^{(m)}_{n+p}$, there exist $\Theta^{(n)}_j$ verifying (ii).

$(ii) \Rightarrow (i)$ Let $v = u\alpha I$. Then,

$$\left\langle x^k P'_{n+1}, v \right\rangle = \sum_{j=-s}^{p} \Theta^{(n)}_j \left\langle x^k P_{n+j}, u \right\rangle = \theta, \quad 0 \leq k \leq n-s-1, \quad n \geq s,$$

and, since $\langle x^{n-s} P_{n-s}, u \rangle$ is non-singular,

$$\left\langle x^{n-s} P'_{n+1}, v \right\rangle = \Theta^{(n)}_{-s} \left\langle x^{n-s} P_{n-s}, u \right\rangle \neq \theta, \quad \text{for some } n \geq s.$$

Thus, $\left(P'_{n+1}\right)_{n\geq 0}$ is quasi-orthogonal of order s with respect to v. Obviously $(P_n)_{n\geq 0}$ is regularly quasi-orthogonal of order p with respect to v and, therefore, u is semi-classical. □

Remark The proof shows that if u is semi-classical with a Pearson-type equation $D(u\alpha I) = uB$, then there is a structure relation with integers $p = \deg \alpha$ and $s = \max\{p-2, q-1\}$, where $q = \deg B$.

5.2 Theorem Differo – differential equation

Let $u \in \mathbb{P}^{(m)'}$ be quasi-definite and let $(P_n)_{n\geq 0}$ be the corresponding sequence of left orthogonal matrix polynomials. Then, the following statements are equivalent:

(i) u is semi-classical.

(ii) There exist two polynomials α, $\beta \in \mathbb{P}$, $\alpha \neq 0$, and matrices $\Lambda^{(n)}_k \in \mathbb{C}^{(m,m)}$ $(n \geq 0, -s \leq k \leq s)$, such that

$$\alpha(x) P''_n(x) + \beta(x) P'_n(x) = \sum_{k=-s}^{s} \Lambda^{(n)}_k P_{n+k}(x),$$

where $s \geq \max\{p-2, q-1\}$, $p = \deg \alpha$ and $q = \deg \beta$ (we use the convention $P_k = \theta$ for $k < 0$).

Proof

$(i) \Rightarrow (ii)$ Let u be semi-classical. By Theorem 5.1, there exist a polynomial $a \in \mathbb{P}$, with $\deg a = p_1 \geq 0$, a non-negative integer $s_1 \geq p_1 - 2$ and matrices $\Theta^{(n)}_j \in \mathbb{C}^{(m,m)}$ $(n \geq 0, \ -s_1 \leq j \leq p_1)$, such that

$$a(x) P'_n(x) = \sum_{j=-s_1}^{p_1} \Theta^{(n-1)}_j P_{n-1+j}(x). \tag{5.1}$$

Taking derivatives in (5.1), we obtain

$$a^2P_n'' + aa'P_n' = \sum_{j=-s_1}^{p_1} \Theta_j^{(n-1)} aP_{n-1+j}'.$$

Using (5.1) in the right hand side of the above equality,

$$a^2P_n'' + aa'P_n' = \sum_{j=-s_1}^{p_1} \Theta_j^{(n-1)} \sum_{k=-s_1}^{p_1} \Theta_k^{(n-2+j)} P_{n-2+j+k},$$

Let us denote $\alpha = a^2$, $\beta = aa'$, $p = 2p_1 = \deg\alpha$, $q = 2p_1 - 1 = \deg\beta$, $s = \max\{2p_1 - 2, 2s_1 + 2\}$ and $\Lambda_k^{(n)} = \sum_{j=-s_1}^{p_1} \Theta_j^{(n-1)}\Theta_{k-j+2}^{(n-2+j)}$. Then,

$$\alpha P_n'' + \beta P_n' = \sum_{j=-s}^{s} \Lambda_j^{(n)} P_{n+j},$$

with $s \geq \max\{p-2, q-1\}$.

$(ii) \Rightarrow (i)$ Without loss of generality, assume that $(P_n)_{n\geq 0}$ is the sequence of monic left orthogonal matrix polynomials with respect to u, that is, $P_n(x) = x^n I + \cdots$. Denoting by $(\Pi_n)_{n\geq 0}$ the sequence of monic right orthogonal matrix polynomials with respect to u, the dual basis $(\pi_n)_{n\geq 0}$ of $(P_n)_{n\geq 0}$ is given by $\pi_n = u\Pi_n E_n^{-1}$, where $E_n := \langle x^n P_n, u\rangle$. Let $(Q_n)_{n\geq 0}$ be the basis of $\mathbb{P}^{(m)}$ given by $Q_n = P_{n+1}'$, and let $(\rho_n)_{n\geq 0}$ be the corresponding dual basis. We have the relation

$$D\rho_n = -\pi_{n+1}, \qquad n \geq 0. \tag{5.2}$$

Consider for every $n \geq 0$ the linear functional $-D(\pi_n\alpha) + \pi_n\beta$. There exists $\left(\lambda_k^{(n)}\right)_{k\geq 0} \subseteq \mathbb{C}^{(m,m)}$ such that

$$-D(\pi_n\alpha) + \pi_n\beta = \sum_{k=0}^{\infty} \rho_k \lambda_k^{(n)},$$

with

$$\left\langle Q_j, -D(\pi_n\alpha) + \pi_n\beta\right\rangle = \sum_{k=0}^{\infty} \left\langle Q_j, \rho_k\right\rangle \lambda_k^{(n)} = \lambda_j^{(n)}, \qquad j \geq 0.$$

Hence, our hypothesis implies that

$$\lambda_j^{(n)} = \sum_{k=-s}^{s} \Lambda_k^{(j+1)} \delta_{j+1+k,n}. \tag{5.3}$$

For $n = 0$, (5.3) leads to $\lambda_j^{(0)} = 0$ if $j \geq s$, and

$$-D(\pi_0\alpha) + \pi_0\beta = \sum_{k=0}^{s-1} \rho_k \lambda_k^{(0)},$$

or, equivalently,

$$-D(u\alpha)E_0^{-1} + u\beta E_0^{-1} = \sum_{k=0}^{s-1} \rho_k \lambda_k^{(0)}.$$

If we denote $\widetilde{\lambda}_k^{(0)} = \lambda_k^{(0)} E_0$, then

$$-D(u\alpha) + u\beta = \sum_{k=0}^{s-1} \rho_k \widetilde{\lambda}_k^{(0)}. \tag{5.4}$$

In a similar way, for $n = 1$, (5.3) gives

$$-D(u\alpha\Pi_1)E_1^{-1} + u\beta\Pi_1 E_1^{-1} = \sum_{k=0}^{s} \rho_k \lambda_k^{(1)},$$

because $\lambda_j^{(1)} = 0$ for $j \geq s+1$. Thus,

$$-D(u\alpha\Pi_1) + u\beta\Pi_1 = \sum_{k=0}^{s} \rho_k \widetilde{\lambda}_k^{(1)}, \tag{5.5}$$

with $\widetilde{\lambda}_k^{(1)} = \lambda_k^{(1)} E_1$.

Keeping in mind that $P_1(x) = xI + M$, (5.5) can be written as

$$-D(u\alpha)\Pi_1 - u\alpha I + u\beta\Pi_1 = \sum_{k=0}^{s} \rho_k \widetilde{\lambda}_k^{(1)},$$

and, using (5.4), we get

$$-u\alpha I = \sum_{k=0}^{s} \rho_k \widetilde{\lambda}_k^{(1)} - \left(\sum_{k=0}^{s-1} \rho_k \widetilde{\lambda}_k^{(0)} \right) \Pi_1.$$

Taking derivatives in the above equality and applying (5.2) we obtain

$$\begin{aligned} D(u\alpha I) &= \sum_{k=0}^{s} \pi_{k+1} \widetilde{\lambda}_k^{(1)} - \left(\sum_{k=0}^{s-1} \pi_{k+1} \widetilde{\lambda}_k^{(0)} \right) \Pi_1 + \sum_{k=0}^{s-1} \rho_k \widetilde{\lambda}_k^{(0)} = \\ &= \sum_{k=0}^{s} u\Pi_{k+1} E_{k+1}^{-1} \widetilde{\lambda}_k^{(1)} - \left(\sum_{k=0}^{s-1} u\Pi_{k+1} E_{k+1}^{-1} \widetilde{\lambda}_k^{(0)} \right) \Pi_1 + \sum_{k=0}^{s-1} \rho_k \widetilde{\lambda}_k^{(0)}. \end{aligned}$$

As a consequence, the polynomial $\Psi \in \mathbb{P}^{(m)}$ given by

$$\Psi := \sum_{k=0}^{s} \Pi_{k+1} E_{k+1}^{-1} \widetilde{\lambda}_k^{(1)} - \left(\sum_{k=0}^{s-1} \Pi_{k+1} E_{k+1}^{-1} \widetilde{\lambda}_k^{(0)} \right) \Pi_1,$$

whose degree is not greater than $s+1$, verifies that

$$D(u\alpha I) = u\Psi + \sum_{k=0}^{s-1} \rho_k \widetilde{\lambda}_k^{(0)}.$$

From this and (5.4) we get

$$D(u\alpha I) = uB$$

with $B = \dfrac{\Psi + \beta I}{2}$, that is, u is semi-classical. □

Remarks

(i) The distributional equation $D(u\alpha) = u\beta$ for a scalar functional u implies that the corresponding orthogonal polynomial sequence (p_n) satisfies

$$\begin{aligned} \langle x^k(\alpha p_n'' + \beta p_n'), u\rangle &= \langle (x^k p_n')', u\alpha\rangle - k\langle x^{k-1} p_n', u\alpha\rangle + \langle x^k p_n', u\beta\rangle = \\ &= -k\langle x^{k-1} p_n', u\alpha\rangle, \end{aligned}$$

which vanishes for $k = 0, \ldots, n-s-1$. So,

$$\alpha p_n'' + \beta p_n' = \sum_{k=-s}^{s} \lambda_{nk} p_{n+k}.$$

In the matricial case, the equality $D(u\alpha I) = uB$ does not imply $\langle x^k P_n', uB\rangle = \langle x^k B P_n', u\rangle$ due to the non-commutativity of the matrix product, and, more generally, it is not possible to ensure the existence of matrix polynomials B_n such that $\langle x^k P_n', uB\rangle = \langle x^k B_n P_n', u\rangle$ neither. Therefore, we can not obtain in this way a differo-differential equation with the polynomial B explicitly.

(ii) Obviuosly, the differo-differential equation given by Theorem 5.2 is not unique because we can modify it by adding any structure relation.

(iii) The proof of Theorem 5.2 shows that, if u is a semi-classical functional with a Pearson-type equation $D(uaI) = uB$, $\deg a = p_1$, $\deg B = q_1$, then

$$\alpha P_n'' + \beta P_n' = \sum_{j=-s}^{p} \Lambda_j^{(n)} P_{n+j},$$

where $p = 2p_1 - 2$, and $s = \max\{2p_1 - 2, 2s_1 + 2\} = \{2p_1 - 2, 2q_1\}$, according to the remark of Theorem 5.1.

6 Some examples

We will illustrate the preceding results with some examples.

1. In the first one, we consider the scalar Laguerre weight $w = xe^{-x}$ that verifies the Pearson-type equation

$$(wx)' = w(2 - x).$$

Now, we define the positive definite matrix weight function

$$u := w \begin{pmatrix} 1 & 1 \\ 1 & 1 + x^2 \end{pmatrix},$$

that also verifies a Pearson-type equation. In fact,

$$D(uxI) = uB(x), \qquad B(x) = \begin{pmatrix} 2 - x & -2 \\ 0 & 4 - x \end{pmatrix}.$$

Moreover,

$$\hat{u} = TuT^t = \begin{pmatrix} xe^{-x} & 0 \\ 0 & x^3e^{-x} \end{pmatrix}, \qquad T = \begin{pmatrix} 1 & 0 \\ -1 & 1 \end{pmatrix},$$

i.e., u is congruent with a diagonal weight with Laguerre classical scalar weights in its diagonal. Let $l_n^{(a)}$ be the monic Laguerre polynomials asociated with the weight $x^a e^{-x}$, $a > -1$. Then,

$$L_n(x) = \begin{pmatrix} l_n^{(1)}(x) & 0 \\ 0 & l_n^{(3)}(x) \end{pmatrix}$$

constitutes a sequence of matrix orthogonal polynomials related to the diagonal weight $\hat{u}$. Obviously $\hat{u}$ satisfies a Pearson-type equation,

$$D(\hat{u}xI) = \hat{u}\hat{B}(x), \qquad \hat{B}(x) = (T^t)^{-1}B(x)T^t = \begin{pmatrix} 2 - x & 0 \\ 0 & 4 - x \end{pmatrix},$$

and a sequence of matrix orthogonal polynomials associated with u is given by

$$P_n(x) = L_n(x)T, \qquad n \geq 0.$$

It is easy to obtain the structure relation

$$xP'_{n+1}(x) = (n+1)P_{n+1} + (n+1)\begin{pmatrix} n+2 & 0 \\ 0 & n+4 \end{pmatrix} P_n(x),$$

as well as, for all $n \geq 2$, a differo-differential equation

$$\begin{aligned} x^2P''_n(x) + xP'_n(x) = n(n+1)P_n(x) + \\ + n^2\begin{pmatrix} n+1 & 0 \\ 0 & n+3 \end{pmatrix} P_{n-1}(x) + \\ + n(n-1)\begin{pmatrix} n(n+1) & 0 \\ 0 & (n+2)(n+3) \end{pmatrix} P_{n-2}(x). \end{aligned}$$

Moreover, the polynomials $l_n^{(a)}$ satisfy the differential equation

$$xl_n^{(a)''}(x) + (x-a-1)l_n^{(a)'}(x) = nl_n^{(a)}(x),$$

and so, we can obtain for the polynomials P_n the following differential equation

$$xP''_n(x) + \begin{pmatrix} x-2 & 0 \\ 0 & x-4 \end{pmatrix} P'_n(x) = nP_n(x).$$

For the monic polynomials $\widetilde{P}_n = T^{-1}P_n$ we have

$$x\widetilde{P}''_n(x) + T^{-1}\begin{pmatrix} x-2 & 0 \\ 0 & x-4 \end{pmatrix} T\widetilde{P}'_n(x) = n\widetilde{P}_n(x),$$

and then

$$x\widetilde{P}''_n(x) + B^t(x)\widetilde{P}'_n(x) = n\widetilde{P}_n(x).$$

The polynomials $\widetilde{P}_n$ can be obtained by means of a Rodrigues-type formula

$$\widetilde{P}_n(x) = \frac{1}{n!}\left(T^tT\right)^{-1} u^{-1}\frac{d^n\left(x^nu\right)}{dx^n}\left(T^tT\right).$$

2. Now, we consider the matrix weight

$$u := e^{-x^2}R(x), \qquad R(x) = \begin{pmatrix} 0 & 1 \\ 1 & x \end{pmatrix}.$$

Notice that u is hermitian but not positive definite. u verifies the distributional equation

$$u' = e^{-x^2}R'(x) - 2xe^{-x^2}R(x) = u(-2xI + R^{-1}(x)R'(x)) = uB(x), \qquad (6.1)$$

where $B(x) = \begin{pmatrix} -2x & 1 \\ 0 & -2x \end{pmatrix}$. Moreover, u has no congruent diagonal weight.

Let

$$P_n(x) = (-2)^{(-n)} e^{x^2} S^{-1}(x) \frac{d^n(e^{-x^2} S(x))}{dx^n}, \qquad n \geq 0, \tag{6.2}$$

where $S(x) = \begin{pmatrix} 1+x & 1 \\ x & 1 \end{pmatrix}$. Derivating (6.2) we get

$$P_{n+1}(x) = -\frac{1}{2}\left(-2xI + S^{-1}(x)S'(x)\right) P_n(x) - \frac{1}{2}P_n'(x),$$

so,

$$P_{n+1}(x) = -\frac{1}{2}\left(B^t(x)P_n(x) + P_n'(x)\right). \tag{6.3}$$

This shows that P_n is a monic matrix polynomial of degree n because $P_0(x) = I$.

Applying the Leibnitz rule for the derivatives in (6.2), we have

$$\begin{aligned} P_{n+2}(x) &= (-2)^{-n-2} e^{x^2} S^{-1}(x) \frac{d^{n+1}(-2xe^{-x^2} S(x) + e^{-x^2} S'(x))}{dx^{n+1}} = \\ &= (-2)^{-n-1} x e^{x^2} S^{-1}(x) \frac{d^{n+1}(e^{-x^2} S(x))}{dx^{n+1}} + \\ &\quad + (n+1)(-2)^{-n-1} e^{x^2} S^{-1}(x) \frac{d^n(e^{-x^2} S(x))}{dx^n} + \\ &\quad + (-2)^{-n-2} e^{x^2} S^{-1}(x) S'(x) S^{-1}(x) \frac{d^{n+1}(e^{-x^2} S(x))}{dx^{n+1}}. \end{aligned}$$

So, we obtain the recurrence relation

$$P_{n+2}(x) = -\frac{1}{2}\left(B^t(x)P_{n+1}(x) + (n+1)P_n(x)\right). \tag{6.4}$$

As a consequence, (P_n) is a sequence of left orthogonal matrix polynomials with respect to certain non-singular functional. Notice that (6.4) implies that the matrix coefficients $E_n = \langle x^n P_n, u\rangle = \langle P_n, P_n\rangle_u$ are given by

$$E_n = \frac{(-1)^n}{2^n} n! E_0,$$

so, the orthogonality matrix functional can not be positive definite.

Moreover, from (6.3) and (6.4) we can obtain the structure relation

$$P_n'(x) = nP_{n-1}(x), \tag{6.5}$$

and, by derivation, the differo-differential equation

$$P_n''(x) = n(n-1)P_{n-2}(x). \tag{6.6}$$

So, (P_n) is a semi-classical matrix orthogonal polynomial sequence. Besides, from (6.3) and (6.5) it follows the differential equation

$$P_n''(x) + B^t(x)P_n'(x) = -2nP_n(x). \tag{6.7}$$

Indeed, (P_n) is a sequence of left orthogonal matrix polynomials with respect to the matrix weight u. In fact,

$$\begin{aligned}\int_{-\infty}^{\infty} P_n'(x)e^{-x^2}R(x)P_k'^t(x)\,dx &= \\ &= \int_{-\infty}^{\infty}\big(P_n''(x)+(-2xI + R'(x)R^{-1}(x))P_n'(x)\big)e^{-x^2}R(x)P_k^t(x)\,dx = \\ &= -2n\int_{-\infty}^{\infty} P_n(x)e^{-x^2}R(x)P_k^t(x)\,dx,\end{aligned}$$

the last equality obtained from equation (6.7). In the same way, we get

$$\int_{-\infty}^{\infty} P_n'(x)e^{-x^2}R(x)P_k'^t(x)dx = -2k\int_{-\infty}^{\infty} P_n(x)e^{-x^2}R(x)P_k^t(x)dx.$$

So, we have the orthogonality of (P_n) with respect to u.

Summarizing, u is a quasi-definite functional such that:

(i) It satisfies the distributional equation $Du = uB$.

(ii) The corresponding sequence of left orthogonal matrix polynomials (P_n) are given by the Rodrigues-type formula (6.2).

(iii) (P_n) verifies the structure relation (6.5), the differo-differential equation (6.6) and the differential equation (6.7). Keeping in mind Remark (i) to Theorem 5.2, notice that we have (6.7) because $uB = (uB)^t = B^t u$ and $B^t P_n = P_n B^t$.

3. In the last example, we consider the Jacobi scalar weight $w = 1 - x^2$ that satisfies the distributional equation

$$\big(w(1-x^2)\big)' = -w4x, \qquad x \in [-1,1].$$

We define the matrix weight function,

$$u := w\begin{pmatrix} 1 & x \\ x & 2-x^2 \end{pmatrix}, \qquad x \in [-1,1],$$

that verifies the Pearson-type equation

$$D\left(u(1-x^2)I\right) = uB(x), \qquad B(x) = \begin{pmatrix} -9x & 2+x^2 \\ 1 & -11x \end{pmatrix}.$$

Notice that u is positive definite and then, there exists a sequence of left orthogonal matrix polynomials (P_n) related to u. Moreover, u is not congruent with a diagonal weight. So, the sequence (P_n) satisfies a structure relation with $s = 1$, $p = 2$, and a differo-differential equation of second order with the form

$$(1-x^2)^2 P_n''(x) - x(1-2x)P_n'(x) = \sum_{k=-4}^{2} \Lambda_{nk} P_{n+k}.$$

Acknowledgements This research was supported by Dirección General de Enseñanza Superior of Spain (DGES), Project PB98-1615.

The authors are very grateful to Professor Francisco Marcellán for his remarks and useful suggestions.

References

[1] M. Alfaro, A. Branquinho, F. Marcellán, J. Petronilho, A generalization of a theorem of S. Bochner, *Publicaciones del Seminario Matemático García de Galdeano*, Serie II, Sección 1, número 11 (1992).

[2] S. Bonan, D.S. Lubinski, P. Nevai, Orthogonal polynomials and their derivatives, SIAM *J. Math. Anal.* **18** (1987) 1163-1175.

[3] T.S. Chihara, *"An introduction to orthogonal polynomials"*, Gordon and Breach, New York, 1978.

[4] A.J. Durán, On orthogonal polynomials with respect to a positive definite matrix of measures, *Can. J. Math.* **47** (1995) 88-112.

[5] A.J. Durán, Matrix inner product having a matrix symmetric second order differential operator, *Rocky Mountain J. Math.* **27** (1997) 585-600.

[6] A.J. Durán, W. Van Assche, Orthogonal matrix polynomials and higher-order recurrence relations, *Linear Alg. Appl.* **219** (1995) 261-280.

[7] J. Favard, Sur les polynômes de Tchebicheff, *C. R. Acad. Sci. Paris* **200** (1935) 2052-2053.

[8] J.S. Geronimo, Scattering theory and matrix orthogonal polynomials on the real line, *Circuits Systems Signal Process* **1** (1982) 471-495.

[9] E. Hendriksen, H. van Rossum, Semi-classical orthogonal polynomials, C.Brezinski et al. Eds., *Lecture Notes in Math.* **1171** (Springer, Berlin, 1985) 354-361.

[10] F. Marcellán, H.O. Yakhlef, Recent trends on analytic properties of matrix orthonormal polynomials, *Electr. Trans. Numer. Anal.* **14** (2002) 127-141.

[11] P. Maroni, Variations around classical orthogonal polynomials. Connected problems, *J. Comput. Appl. Math.* **48** (1-2) (1993) 133-155.

[12] P. Maroni, Une théorie algébrique des polynômes orthogonaux. Application aux polynômes orthogonaux semiclassiques. C. Brezinski et al. Eds. *Orthogonal Polynomials and Their Applications, IMACS Ann. Comput. Appl. Math.* **9** (1991) 95-130.

[13] G. Sansigre, *Polinomios ortogonales matriciales y matrices bloques*, Doctoral Dissertation, Universidad de Zaragoza (1992). In Spanish.

[14] J.A. Shohat, A differential equation for orthogonal polynomials, *Duke. Math. J.* **5** (1939) 401-407.

JOURNAL OF CONCRETE AND APPLICABLE MATHEMATICS,VOL.3,NO.3,335-345,2005,COPYRIGHT 2005 EUDOXUS PRESS,LLC

An Euler-type quadrature rule derived by using Radon's method

by **Carlo Belingeri** and **Gabriella Bretti**

Rome University "La Sapienza"

Abstract

The use of Euler polynomials and Euler numbers allows us to construct a quadrature rule similar to the well known Euler-MacLaurin quadrature formula.

1991 Mathematics Subject Classification. 33C45, 33F05, 65D32.

Key words and phrases. Euler polynomials, Euler numbers, Euler-MacLaurin quadrature formula, Quadrature rules, Differential operators.

1 Introduction

Among the composite quadrature rules the Euler-MacLaurin formula is of particular importance since the extrapolation methods proposed by Romberg are essentially based on it. The Romberg formulas constitute one of the most efficient methods to estimate integrals (see [5]).

It has been recently proved thet generalized quadrature procedures can be constructed employing Appell polynomials [4] and containing as a particular case either the Euler-MacLaurin formula or a similar formula using the Euler numbers (instead of the Bernoulli ones) and even derivatives (instead od odd) evaluated at the extrema of the considered integral.

As it is well known, in [1] can be found a general theory of quadrature rules which is founded on a method going back to Radon, see [2], reducing any quadrature rule known in literature to evaluate the integral of $g(x)u(x)$ ($g(x)$ weight function) to the construction of a suitable differential operator E, and of particular solutions of the adjoint equation $E^*(\phi) = g$.

It is proved in [1] how the above mentioned operators can be chosed in order to recover the Gaussian quadrature rules and the Euler-MacLaurin formula too.

In this paper we extend the Radon technique to the quadrature rule considered in [3], whereas in [3] this formula is derived by using integration by parts. This new approach shows that the Radon method works even in this case.

2 Elementary quadrature rule in a finite interval

In [1], chapter 2, an elementary quadrature formula which is valid for real-valued functions defined in a finite interval is constructed. The method can be summarized as follows.

Let be $g(x) \in L[a,b]$ the weight, $u(x) \in AC^{n-1}[a,b]$ the argument function and denote by $a = x_0 < x_1 < \ldots < x_m < x_{m+1} = b$ the nodes of the quadrature rule:

$$\int_a^b g(x)u(x)dx = \sum_{h=0}^{n-1}\sum_{i=1}^{m} A_{hi}u^{(h)}(x_i) + R(u) \tag{2.1}$$

associated to a linear differential operator E of order n

$$E(u) = \sum_{k=0}^{n} a_k(x)\frac{d^{n-k}}{dx^{n-k}}u. \tag{2.2}$$

We deal with the problem of finding the coefficients A_{hi} in equation (2.1) in order to satisfy the condition:

$$E(u) = 0 \Rightarrow R(u) = 0. \tag{2.3}$$

Let be

$$E_r = \sum_{k=0}^{r} a_k(x)\frac{d^{r-k}}{dx^{r-k}}, \quad r = 0, 1, \ldots, n-1$$

the reduced operators associated with the operator E, and let be

$$E^*(v) = \sum_{k=0}^{n}(-1)^{n-k}\frac{d^{n-k}}{dx^{n-k}}[a_k(x)v(x)] \qquad (a_0(x) = 1), \tag{2.4}$$

the adjoint operator of E.

Then the following theorem is valid:

Theorem 1. Denoting by E_r^* the adjoint operators of the reduced operators E_r, a necessary condition in order that (2.3) holds true is to assume:

$$A_{hi} = [E^*_{n-h-1}(\varphi_i(x) - \varphi_{i-1}(x))]_{x=x_i} \tag{2.5}$$

where $\varphi_1(x), \ldots, \varphi_{m-1}(x)$ are arbitrary solutions of the differential equation

$$E^*(\varphi) = g(x) \tag{2.6}$$

while $\varphi_0(x)$ and $\varphi_m(x)$ are integrals of the same equation (2.6) satisfying the initial conditions

$$\varphi_0^{(h)}(a) = 0, \ \varphi_m^{(h)}(a) = 0, \quad h = 0, 1, \ldots, n-1. \tag{2.7}$$

As a consequence the remainder $R(u)$ can be written in the form:

$$R(u) = \int_a^b \phi(x)E(u)dx = \sum_{i=0}^{m} \int_{x_i}^{x_{i+1}} \varphi_i(x)E(u)dx, \tag{2.8}$$

where $\phi(x) = \varphi_i(x), \ \ x \in (x_i, x_{i+1}), \ \ i = 0, 1, \ldots, m-1, m.$

It is well known that if the function $\phi(x)$ is continuous and does not change sign in $[a, b]$, then the estimate of the remainder $R(u)$ can be simplified (see [1]).

3 The Euler-MacLaurin quadrature formula

In [1], chapter 4, the Euler-MacLaurin formula is constructed using Bernoulli numbers. A synthesis of the method is as follows.

Let us assume

$$a = 0, \quad b = 1, \quad g(x) = 1, \quad x_1 = 0, \quad x_2 = 1,$$

$$m = 2, \quad n = 2\nu + 2, \quad E = \frac{d^{2\nu+2}}{dx^{2\nu+2}}, \quad \nu = 0, 1, \ldots$$

$$\varphi_1(x) = \frac{B_{2\nu+2}(x) - \mathcal{B}_{2\nu+2}}{(2\nu+2)!},$$

where $B_{2\nu+2}(x)$ and $\mathcal{B}_{2\nu+2}$ are respectively Bernoulli polynomials and Bernoulli numbers. The coefficients in (2.1) are given by

$$A_{h1} = \begin{cases} 0 & \text{h even} \\ \frac{\mathcal{B}_{h+1}}{(h+1)!} & \text{h odd} \end{cases}$$

$$A_{h2} = \begin{cases} 0 & \text{h even} \\ -\frac{\mathcal{B}_{h+1}}{(h+1)!} & \text{h odd} \end{cases}$$

$h = 1, \ldots, 2\nu$,

$$A_{0,1} = A_{0,2} = \frac{1}{2} \; ; \; A_{2\nu+1,1} = A_{2\nu+1,2} = 0.$$

Then (2.1) becomes

$$\int_0^1 u(x)dx = \frac{1}{2}[u(0) + u(1)] + \sum_{k=1}^{\nu} \frac{\mathcal{B}_{2k}}{(2k)!}[u^{2k-1}(0) - u^{2k-1}(1)] + R(u)$$

and (2.8), under the hypotesis $u^{2\nu+2}(x) \in C[0,1]$, becomes

$$R(u) = \int_0^1 \frac{B_{2\nu+2}(x) - \mathcal{B}_{2\nu+2}}{(2\nu+2)!} u^{(2\nu+2)}(x)dx = -\frac{B_{2\nu+2}(x)}{(2\nu+2)!} u^{(2\nu+2)}(\xi), \quad 0 < \xi < 1.$$

4 Euler polynomials and Euler numbers: definition and properties

The Euler polynomials $E_n(x)$ are defined by the following generating function:

$$\frac{2e^{xt}}{e^t + 1} = \sum_{n=0}^{\infty} E_n(x) \frac{t^n}{n!}, \qquad |t| < \pi; \tag{4.1}$$

the Euler numbers $\mathcal{E}_n$ can be obtained by the generating function:

$$\frac{2}{e^t + e^{-t}} = \sum_{n=0}^{\infty} \frac{\mathcal{E}_n}{n!} t^n, \tag{4.2}$$

and the connection between Euler numbers and Euler polynomials is given by

$$E_n\left(\frac{1}{2}\right) = 2^{-n}\mathcal{E}_n, \qquad n = 0, 1, 2, \ldots \; . \tag{4.3}$$

A recursive computation of the Euler polynomials can be obtained by using the following formula:

$$E_n(x) + \sum_{k=0}^{n} \binom{n}{k} E_k(x) = 2x^n, \qquad n = 1, 2, \ldots \; . \tag{4.4}$$

The first Euler polynomials are given by

$$E_0(x) = 1,\ E_1(x) = x - \frac{1}{2},\ E_2(x) = x^2 - x,\ E_3(x) = x^3 - \frac{3}{2}x^2 + \frac{1}{4}\ , \ldots$$

and the first Euler numbers are given by

$$\mathcal{E}_0 = 1,\ \mathcal{E}_1 = 0,\ \mathcal{E}_2 = -1,\ \mathcal{E}_3 = 0,\ \mathcal{E}_4 = 5,\ \mathcal{E}_5 = 0,\ \mathcal{E}_6 = -61,\ \mathcal{E}_7 = 0,\ \ldots\ .$$

For further values see [8], p. 810.

The Euler polynomials satisfy the following relations:

$$E_n^k(x) = \frac{n!}{(n-k)!} E_{n-k}(x), \qquad k = 0, 1, \ldots, n \tag{4.5}$$

$$\int_a^x E_n(t)dt = \frac{E_{n+1}(x) - E_{n+1}(a)}{n+1} \tag{4.6}$$

$$E_n(x) = \sum_{k=0}^{n} \binom{n}{k} \frac{\mathcal{E}_k}{2^k} \left(x - \frac{1}{2}\right)^{n-k}. \tag{4.7}$$

For $x = 0$ and $x = 1$, by (4.7), can be obtained $E_n(0)$ and $E_n(1)$, used in the following; the first values are given by

$$E_0(0) = 1,\ \ E_1(0) = -\frac{1}{2},\ \ E_2(0) = 0,\ \ E_3(0) = \frac{1}{4},\ \ E_4(0) = 0,\ \ E_5(0) = -\frac{1}{2}, \ldots$$

$$E_0(1) = 1,\ \ E_1(1) = \frac{1}{2},\ \ E_2(1) = 0,\ \ E_3(1) = -\frac{1}{4},\ \ E_4(1) = 0,\ \ E_5(1) = \frac{1}{2}, \ldots$$

5 The Euler-type quadrature formula

In this section, starting from the Euler polynomials, we construct a quadrature rule similar to the well known Euler-MacLaurin quadrature formula, but using Euler (instead of Bernoulli) numbers in the remainder term, and even (instead

of odd) order derivatives of the given function evalueted at the extrema of the considered interval. In a forthcoming paper the same procedure will be generalized starting from the Appell polynomials.

Let us assume

$$a = 0, \quad b = 1, \quad g(x) = 1, \quad x_1 = 0, \quad x_2 = 1,$$

$$m = 2, \quad n = 2\nu + 2, \quad E = \frac{d^{2\nu+2}}{dx^{2\nu+2}}, \quad \nu = 0, 1, \ldots .$$

The differential equation $E^*(\varphi) = g(x)$ is to be written $\varphi^{(2\nu+2)} = 1$. We need not to consider the solutions $\varphi_0(x), \varphi_2(x)$ and it is sufficient to fix arbitrarily the solution $\varphi_1(x)$; for instance we may assume

$$\varphi_1(x) = \frac{E_{2\nu+2}(x)}{(2\nu+2)!} \tag{5.1}$$

where $E_{2\nu+2}(x)$ denote the Euler polynomial of index $2\nu + 2$.

Equations (2.1) and (2.8) become, under the hypotesis $u(x) \in AC^{2\nu+1}[0, 1]$:

$$\int_0^1 u(x)dx = \sum_{h=0}^{2\nu+1} [A_{h1}u^{(h)}(0) + A_{h2}u^{(h)}(1)] + R(u) \tag{5.2}$$

$$R(u) = \int_0^1 \frac{E_{2\nu+2}(x)}{(2\nu+2)!} u^{(2\nu+2)}(x)dx \tag{5.3}$$

Now we have only to compute the coefficients A_{h1}, A_{h2} using (2.5). We find

$$A_{h1} = [E^*_{2\nu+1-h}(\varphi_1)]_{x=0} = (-1)^{2\nu+1-h}\varphi_1^{(2\nu+1-h)}(0) =$$

$$= \begin{cases} (-1)^{h+1}\frac{E_{2\nu+2}^{(2\nu+1-h)}(0)}{(2\nu+2)!} & h = 0, 1, \ldots, 2\nu \\ \\ \frac{E_{2\nu+2}(0)}{(2\nu+2)!} & h = 2\nu + 1 \end{cases}$$

$$A_{h2} = [E^*_{2\nu+1-h}(-\varphi_1)]_{x=1} = -(-1)^{2\nu+1-h}\varphi_1^{(2\nu+1-h)}(1) =$$

$$= \begin{cases} (-1)^{h+1}\frac{E_{2\nu+2}^{(2\nu+1-h)}(1)}{(2\nu+2)!} & h = 0, 1, \ldots, 2\nu \\ \\ -\frac{E_{2\nu+2}(1)}{(2\nu+2)!} & h = 2\nu + 1 \end{cases}$$

and therefore, recalling (4.5) and (4.7):

$$A_{0,1} = \frac{1}{2}, \quad A_{0,2} = \frac{1}{2}$$

$$A_{h1} = \begin{cases} 0 & \text{h odd} \\ \frac{E_{h+1}(0)}{(h+1)!} & \text{h even} \end{cases}$$

$$A_{h2} = \begin{cases} 0 & \text{h odd} \\ -\frac{E_{h+1}(1)}{(h+1)!} & \text{h even} \end{cases}$$

$h = 1, 2, \ldots, 2\nu$,

$$A_{2\nu+1,1} = 0, \quad A_{2\nu+1,2} = 0.$$

Equation (5.2), recalling the property

$$E_{2s+1}(1) = -E_{2s+1}(0), \quad s = 1, 2, \ldots,$$

becomes

$$\int_0^1 u(x)dx = \frac{1}{2}[u(0) + u(1)] - \sum_{k=1}^{\nu} \frac{E_{2k+1}(0)}{(2k+1)!}[u^{2k}(0) + u^{2k}(1)] + R(u) \qquad (5.4)$$

and (5.3), recalling (4.6), yields

$$R(u) = u^{(2\nu+2)}(\xi) \int_0^1 \frac{E_{2k+2}(x)}{(2k+2)!}dx = \frac{E_{2\nu+3}(1) - E_{2\nu+3}(0)}{(2\nu+3)(2\nu+2)!} u^{(2\nu+2)}(\xi) =$$

$$= -2\frac{E_{2\nu+3}(0)}{(2\nu+3)!} u^{(2\nu+2)}(\xi), \quad 0 < \xi < 1, \quad \nu = 0, 1, \ldots, \qquad (5.5)$$

under the hypotesis $u^{(2\nu+2)}(x) \in C[0,1]$.

Then we have proved the following theorem:

Theorem 2. Let $u(x) \in AC^{2\nu+1}[0,1]$ then the quadrature rule

$$\int_0^1 u(x)dx = \frac{1}{2}[u(0) + u(1)] - \sum_{k=1}^{\nu} \frac{E_{2k+1}(0)}{(2k+1)!}[u^{2k}(0) + u^{2k}(1)] + R(u)$$

holds true, where the remainder $R(u)$ is expressed by

$$R(u) = \int_0^1 \frac{E_{2\nu+2}(x)}{(2\nu+2)!} u^{(2\nu+2)}(x)dx$$

6 Numerical example

Let us apply the quadrature formula (5.4) assuming $u(x) = \frac{sin(x)}{x}$, $\nu = 5$. Then we found

$$\mathcal{I} = \int_0^1 \frac{sin(x)}{x} dx = \frac{1}{2}[u(0) + u(1)] - \sum_{k=1}^{5} \frac{E_{2k+1}(0)}{(2k+1)!}[u^{(2k)}(0) + u^{(2k)}(1)] + R(u),$$

and consequently

$$\mathcal{I} = \frac{1}{2}(1.8414709) - \frac{1}{4 \cdot 3!}(-0.572466) + \frac{1}{2 \cdot 5!}(0.333073) - \frac{17}{8 \cdot 7!}(-0.232804) +$$
$$+ \frac{31}{2 \cdot 9!}(0.277777) - \frac{691}{4 \cdot 11!}(-0.090909) + R(u) = 0.946086418 + R(u),$$

where $R(u) = [-2\frac{-\frac{5461}{2}}{13 \cdot 12!}]u^{(12)}(\xi) = 0.00000087 \cdot u^{(12)}(\xi), \quad 0 < \xi < 1.$

Acknowledgements.

We are grateful to the anonymous referee for pointing out the papers [9], [10], [11], [12], which are connected with our problem, but in the above articles an explicit quadrature rule using Euler polynomials and numbers does not appear.

References

[1] A. Ghizzetti – A. Ossicini: *Quadrature formulae,* Birkhäuser, Basel - Stuttgart, 1970.

[2] J. Radon: *Restausdrücke bei Interpolations und Quadraturformeln durch bestimmte Integrale,* Monatsh. Math., **42** (1935).

[3] G. Bretti – P.E. Ricci: *Euler polynomials and the related quadrature rule,* Georgian Math. J., (to appear).

[4] G. Bretti – M.X. He – P.E. Ricci: *On quadrature rules associated with Appell polynomials,* (submitted).

[5] J. Stoer: *Introduzione all'Analisi Numerica*, Zanichelli, Bologna, 1972.

[6] L. Euler: *Instituziones calculi differentialis*, in Opera Omnia, Series prima: Opera mathematica, **10**, G. Kowalewski Ed., Teubner, 1980.

[7] I.S. Gradshteyn – I.M. Ryzhik: *Table of integrals, series, and products*, Academic Press, New York, 1980.

[8] M. Abramowitz – I.A. Stegun: *Handbook of Mathematical Functions*, Dover, New York, 1965.

[9] Lj. Dedić – M. Matić – J. Pecarić: *On Euler trapezoid formulae*, Appl. Math. & Comp., **123** (2001), 37–62.

[10] Lj. Dedić – M. Matić – J. Pecarić: *Some inequalities of Euler-Grüss type*, Computers Math. Applic., **42** (2001), 843–856.

[11] Lj. Dedić – C.E.M. Pearce – J. Pecarić: *The Euler Formulae and convex functions*, Math. Inequal. Applic., **3** (2000), 211–221.

[12] M. Matić – J. Pecarić – N. Ujecić: *On new estimation of remainder in generalised Taylor's formula*, Math. Inequal. Applic., **2** (1999), 343–361.

Addresses:

Carlo Belingeri

Dipartimento di Metodi e Modelli Matematici

per le Scienze Applicate

Università degli Studi di Roma “La Sapienza”, Italy

Via A. Scarpa, 10

00161 - Roma

Email: belingeri@dmmm.uniroma1.it

Gabriella Bretti

Dipartimento di Matematica “Guido Castelnuovo”

Università degli Studi di Roma “La Sapienza”, Italy

P.le A. Moro, 2

00185 - Roma

Email: bretti@mat.uniroma1.it

JOURNAL OF CONCRETE AND APPLICABLE MATHEMATICS,VOL.3,NO.3,347-352,2005,COPYRIGHT 2005 EUDOXUS PRESS,LLC

A note on quasi power increasing sequences

Hüseyin Bor

Department of Mathematics, Erciyes University, 38039 Kayseri, Turkey

E-mail:bor@erciyes.edu.tr, URL:http://fef.erciyes.edu.tr/math/hbor.htm

Abstract

In this paper a theorem of Bor [3] has been proved under weaker conditions by using a quasi β-power increasing sequence instead of an almost increasing sequence.

2000 AMS Subject Classification: 40D15, 40F05, 40G99.

Key Words: Absolute summability, quasi power increasing sequence.

1 Introduction

A positive sequence (b_n) is said to be almost increasing if there exists a positive increasing sequence (c_n) and two positive constants A and B such that $Ac_n \leq b_n \leq Bc_n$ (see [1]). We denote by $\mathcal{BV}_{\mathcal{O}}$ the $\mathcal{BV} \cap \mathcal{C}_{\mathcal{O}}$, where $\mathcal{C}_{\mathcal{O}}$ and $\mathcal{BV}$ of the null sequences and sequences with bounded variation, respectively. Let $\sum a_n$ be a given infinite series with partial sums (s_n). Let (p_n) be a sequence of positive numbers such that

$$P_n = \sum_{v=0}^{n} p_v \to \infty \quad as \quad n \to \infty, \quad (P_{-i} = p_{-i} = 0, i \geq 1). \tag{1}$$

The sequence-to-sequence transformation

$$U_n = \frac{1}{P_n} \sum_{v=0}^{n} p_v s_v \tag{2}$$

defines the sequence (U_n) of the $(\bar{N}, p_n)$ mean of the sequence (s_n), generated by the sequence of coefficients (p_n) (see [4]).

The series $\sum a_n$ is said to be summable $\mid \bar{N}, p_n \mid_k$, $k \geq 1$, if (see [2])

$$\sum_{n=1}^{\infty} (P_n/p_n)^{k-1} \mid \Delta U_{n-1} \mid^k < \infty \tag{3}$$

where

$$\Delta U_{n-1} = -\frac{p_n}{P_n P_{n-1}} \sum_{v=1}^{n} P_{v-1} a_v, \quad n \geq 1. \tag{4}$$

In the special case when $p_n = 1$ for all values of n, $\mid \bar{N}, p_n \mid_k$ summability is the same as $\mid C, 1 \mid_k$ summability. Also if we take $p_n = \frac{1}{n+1}$,then $\mid \bar{N}, p_n \mid_k$ summability reduces to $\mid \bar{N}, \frac{1}{n+1} \mid_k$.

Quite recently Bor [3] has proved the following theorem.

Theorem A. Let (X_n) be an almost increasing sequence and let there be sequences (β_n) and (λ_n) such that

$$\mid \Delta\lambda_n \mid \leq \beta_n \tag{5}$$

$$\beta_n \to 0 \quad as \quad n \to \infty \tag{6}$$

$$\sum_{n=1}^{\infty} n \mid \Delta\beta_n \mid X_n < \infty \tag{7}$$

$$\mid \lambda_n \mid X_n = O(1) \quad as \quad n \to \infty. \tag{8}$$

If

$$\sum_{n=1}^{m} \frac{\mid \lambda_n \mid}{n} = O(1) \quad as \quad m \to \infty, \tag{9}$$

$$\sum_{n=1}^{m} \frac{1}{n} \mid t_n \mid^k = O(X_m) \quad as \quad m \to \infty, \tag{10}$$

and (p_n) is a sequence such that

$$\sum_{n=1}^{m} \frac{p_n}{P_n} \mid t_n \mid^k = O(X_m) \quad as \quad m \to \infty, \tag{11}$$

where

$$t_n = \frac{1}{n+1} \sum_{v=1}^{n} v a_v, \tag{12}$$

then the series $\sum a_n \lambda_n$ is summable $\mid \bar{N}, p_n \mid_k$ for $k \geq 1$.

2. The main result. The aim of this paper is to prove Theorem A under weaker conditions. For this we need the concept of quasi β-power increasing sequence. A positive sequence (γ_n) is said to be quasi β-power increasing sequence if there exists a constant $K = K(\beta, \gamma) \geq 1$ such that

$$Kn^{\beta}\gamma_n \geq m^{\beta}\gamma_m \tag{13}$$

holds for all $n \geq m \geq 1$. It should be noted that every almost increasing sequence is quasi β-power increasing sequence for any nonnegative β, but the convers need not be true as can be seen by taking the example, say $\gamma_n = n^{-\beta}$ for $\beta > 0$. So we are weakening the hypotheses of the theorem replacing an almost increasing sequence by a quasi β-power increasing sequence.

Now, we shall prove the following theorem:

Theorem. Let $(\lambda_n) \in \mathcal{BV}_{\mathcal{O}}$ and let (X_n) be a quasi β-power increasing sequence for some $0 < \beta < 1$. If all the conditions of Theorem A are satisfied, then the series $\sum a_n\lambda_n$ is summable $\mid \bar{N}, p_n \mid_k$ for $k \geq 1$.

Remark. If we take (X_n) as an almost incrasing sequence, then we get Theorem A. In this case the condition $(\lambda_n) \in \mathcal{BV}_{\mathcal{O}}$ is not needed.

We need the following lemma for the proof of our theorem.

Lemma ([5]). Under the conditions on (X_n), (β_n) and (λ_n) as taken in the statement of the theorem, the following conditions hold, when (7) is satisfied:

$$n\beta_n X_n = O(1) \quad as \quad n \to \infty, \tag{14}$$

$$\sum_{n=1}^{\infty} \beta_n X_n < \infty. \tag{15}$$

3. Proof of the Theorem. Let (T_n) denotes the $(\bar{N}, p_n)$ mean of the series $\sum a_n\lambda_n$. Then, by definition and changing the order of summation, we have

$$T_n = \frac{1}{P_n}\sum_{v=0}^{n} p_v \sum_{i=0}^{v} a_i\lambda_i = \frac{1}{P_n}\sum_{v=0}^{n}(P_n - P_{v-1})a_v\lambda_v.$$

Then, for $n \geq 1$, we have

$$T_n - T_{n-1} = \frac{p_n}{P_nP_{n-1}}\sum_{v=1}^{n} P_{v-1}a_v\lambda_v = \frac{p_n}{P_nP_{n-1}}\sum_{v=1}^{n}\frac{P_{v-1}\lambda_v}{v}va_v.$$

By Abel's transformation, we have

$$\begin{aligned} T_n - T_{n-1} &= \frac{n+1}{nP_n} p_n t_n \lambda_n - \frac{p_n}{P_n P_{n-1}} \sum_{v=1}^{n-1} p_v t_v \lambda_v \frac{v+1}{v} + \frac{p_n}{P_n P_{n-1}} \sum_{v=1}^{n-1} P_v \Delta\lambda_v t_v \frac{v+1}{v} \\ &+ \frac{p_n}{P_n P_{n-1}} \sum_{v=1}^{n-1} P_v t_v \lambda_{v+1} \frac{1}{v} \\ &= T_{n,1} + T_{n,2} + T_{n,3} + T_{n,4}, \quad say. \end{aligned}$$

Since

$$| T_{n,1} + T_{n,2} + T_{n,3} + T_{n,4} |^k \leq 4^k (| T_{n,1} |^k + | T_{n,2} |^k + | T_{n,3} |^k + | T_{n,4} |^k),$$

to complete the proof of the theorem, it is enough to show that

$$\sum_{n=1}^{\infty} (P_n/p_n)^{k-1} | T_{n,r} |^k < \infty \quad for \quad r = 1, 2, 3, 4. \tag{16}$$

Firstly, we have that

$$\begin{aligned} \sum_{n=1}^{m} (P_n/p_n)^{k-1} | T_{n,1} |^k &= O(1) \sum_{n=1}^{m} \frac{p_n}{P_n} | \lambda_n |^{k-1} | \lambda_n | | t_n |^k \\ &= O(1) \sum_{n=1}^{m} | \lambda_n | \frac{p_n}{P_n} | t_n |^k \\ &= O(1) \sum_{n=1}^{m-1} \Delta | \lambda_n | \sum_{v=1}^{n} \frac{p_v}{P_v} | t_v |^k \\ &+ O(1) | \lambda_m | \sum_{n=1}^{m} \frac{p_n}{P_n} | t_n |^k \\ &= O(1) \sum_{n=1}^{m-1} | \Delta\lambda_n | X_n + O(1) | \lambda_m | X_m \\ &= O(1) \sum_{n=1}^{m-1} \beta_n X_n + O(1) | \lambda_m | X_m = O(1) \quad as \quad m \to \infty, \end{aligned}$$

by (5), (8), (11) and (15). Now, when $k > 1$ applying Hölder's inequality with indices k and k', where $\frac{1}{k} + \frac{1}{k'} = 1$, as in $T_{n,1}$, we have that

$$\sum_{n=2}^{m+1} (P_n/p_n)^{k-1} | T_{n,2} |^k = O(1) \sum_{n=2}^{m+1} \frac{p_n}{P_n P_{n-1}} \{ \sum_{v=1}^{n-1} p_v | \lambda_v |^k | t_v |^k \}$$

$$\begin{aligned} &\times \quad \{\frac{1}{P_{n-1}}\sum_{v=1}^{n-1} p_v\}^{k-1} \\ &= \quad O(1)\sum_{v=1}^{m} p_v \mid \lambda_v \mid^{k-1} \mid \lambda_v \mid\mid t_v \mid^k \sum_{n=v+1}^{m+1} \frac{p_n}{P_n P_{n-1}} \\ &= \quad O(1)\sum_{v=1}^{m} \mid \lambda_v \mid \frac{p_v}{P_v} \mid t_v \mid^k = O(1) \quad as \quad m \to \infty. \end{aligned}$$

Again, we have that

$$\begin{aligned} \sum_{n=2}^{m+1}(P_n/p_n)^{k-1} \mid T_{n,3} \mid^k &= O(1)\sum_{n=2}^{m+1}\frac{p_n}{P_n P_{n-1}}\{\sum_{v=1}^{n-1} \mid \Delta\lambda_v \mid P_v \mid t_v \mid^k\} \\ &\times \quad \{\frac{1}{P_{n-1}}\sum_{v=1}^{n-1} P_v \mid \Delta\lambda_v \mid\}^{k-1} \\ &= O(1)\sum_{v=1}^{m}\beta_v P_v \mid t_v \mid^k \sum_{n=v+1}^{m+1}\frac{p_n}{P_n P_{n-1}} \\ &= O(1)\sum_{v=1}^{m}\beta_v \mid t_v \mid^k = O(1)\sum_{v=1}^{m-1} v\beta_v \frac{1}{v} \mid t_v \mid^k \\ &= O(1)\sum_{v=1}^{m-1}\Delta(v\beta_v)\sum_{i=1}^{v}\frac{1}{i} \mid t_i \mid^k + O(1)m\beta_m \sum_{v=1}^{m}\frac{1}{v} \mid t_v \mid^k \\ &= O(1)\sum_{v=1}^{m-1} \mid \Delta(v\beta_v) \mid X_v + O(1)m\beta_m X_m \\ &= O(1)\sum_{v=1}^{m-1} \mid (v+1)\Delta\beta_v - \beta_v \mid X_v + O(1)m\beta_m X_m \\ &= O(1)\sum_{v=1}^{m-1} vX_v \mid \Delta\beta_v \mid + O(1)\sum_{v=1}^{m-1} \mid \beta_v \mid X_v + O(1)m\beta_m X_m \\ &= O(1) \quad as \quad m \to \infty, \end{aligned}$$

by (5), (7), (10), (14) and (15).

Finally, we have that

$$\begin{aligned} \sum_{n=1}^{m}(P_n/p_n)^{k-1} \mid T_{n,4} \mid^k &= O(1)\sum_{n=2}^{m+1}\frac{p_n}{P_n P_{n-1}}\sum_{v=1}^{n-1} P_v \mid \lambda_{v+1} \mid\mid t_v \mid^k \frac{1}{v} \\ &\times \quad \{\frac{1}{P_{n-1}}\sum_{v=1}^{n-1} P_v \frac{\mid \lambda_{v+1} \mid}{v}\}^{k-1} \\ &= O(1)\sum_{v=1}^{m} P_v \mid \lambda_{v+1} \mid\mid t_v \mid^k \frac{1}{v}\sum_{n=v+1}^{m+1}\frac{p_n}{P_n P_{n-1}} \end{aligned}$$

$$
\begin{aligned}
&= O(1)\sum_{v=1}^{m} \mid \lambda_{v+1} \mid \frac{\mid t_v \mid^k}{v} \\
&= O(1)\sum_{v=1}^{m-1} \Delta \mid \lambda_{v+1} \mid \sum_{r=1}^{v} \frac{1}{r} \mid t_r \mid^k + O(1) \mid \lambda_{m+1} \mid \sum_{v=1}^{m} \frac{1}{v} \mid t_v \mid^k \\
&= O(1)\sum_{v=1}^{m-1} \mid \Delta\lambda_{v+1} \mid X_{v+1} + O(1) \mid \lambda_{m+1} \mid X_{m+1} \\
&= O(1)\sum_{v=1}^{m-1} \beta_{v+1} X_{v+1} + O(1) \mid \lambda_{m+1} \mid X_{m+1} \\
&= O(1) \quad as \quad m \to \infty,
\end{aligned}
$$

by (5), (8), (9), (10) and (15). Therefore, we get that

$$\sum_{n=1}^{m} (P_n/p_n)^{k-1} \mid T_{n,r} \mid^k = O(1) \quad as \quad m \to \infty, \quad for \quad r = 1, 2, 3, 4.$$

This completes the proof of the theorem.

If we take $p_n = 1$ for all values of n in this theorem, then we get a new result concerning the $\mid C, 1 \mid_k$ summability factors.

Also, if we take $p_n = \frac{1}{n+1}$ in this theorem, then we get another new result concerning the $\mid \bar{N}, \frac{1}{n+1} \mid_k$ summability factors.

References

[1] S. Aljancic and D. Arandelovic, O-regularly varying functions. Publ. Inst. Math., 22 (1977), 5-22.

[2] H. Bor, A note on two summability methods, Proc. Amer. Math. Soc., 98 (1986), 81-84.

[3] H. Bor, On absolute Riesz summability factors, Adv. Stud. Contemp. Math., 3 (2001), 23-29.

[4] G.H. Hardy, Divergent Series, Oxford University Press, 1949.

[5] L. Leindler, A new application of quasi power increasing sequences, Publ. Math. Debrecen, 58 (2001), 791-796.

JOURNAL OF CONCRETE AND APPLICABLE MATHEMATICS,VOL.3,NO.3,353-376,2005,COPYRIGHT 2005 EUDOXUS PRESS,LLC

Portfolio Choice With Heavy Tailed Distributions

Sergio Ortobelli
University of Bergamo, Italy
Almira Biglova
University of Karlsruhe, Germany
Isabella Huber
University of Karlsruhe, Germany
Borjana Racheva
FinAnalytica Inc., Sofia, Bulgaria
Stoyan Stoyanov
FinAnalytica Inc., Sofia, Bulgaria

ABSTRACT: This paper analyzes portfolio selection models with heavy tailed return distributions. Firstly, we examine investor's optimal choices when we assume respectively either Gaussian or stable non-Gaussian unconditional distributed index returns. Then, we approximate discrete time optimal allocations assuming returns following an ARMA process. Finally, we describe further autoregressive portfolio choice models.
2000 AMS SUBJECT CLASSIFICATION: 91B28, 60E07, 60G10
KEY WORDS: Stable distributions, portfolio selection, ARMA models.

1. INTRODUCTION

Over the last fifty years, the problem of optimal portfolio selection has lost none of its allure or importance for the financial community. Just consider the following reason, which will appeal immediately to every investor. At a time when more than fifty percent of all financial assets in North America are controlled by pension or mutual funds, a lot of people apparently employ someone else to manage their money. Also, they are obviously willing to pay high fees or expenses for these services and are naturally very interested in how well "their" funds are performing.

The mean-variance analysis, developed by Markowitz and Tobin, generalized into an equilibrium theory by Sharpe, Lintner and Mossin and into an inter-temporal theory by Samuelson and Merton, was the first theory to give rigorous results to the portfolio selection problem in terms of the mean and the variance. However, many criticisms and empirical rejections have underlined the intrinsic limits of the mean-variance approximation. Probably Roll in [23], [24], [25] was the first to clearly understand the weaknesses of the theory and the empirical deficiencies. On the other hand, the fundamental work of Mandelbrot [15], [16] and Fama [10] has sparked considerable interest in studying the

empirical distribution of financial assets. The excess kurtosis, found in Mandelbrot's and Fama's investigations, led them to reject the normal assumption (generally used to justify the mean-variance approach) and to propose the stable Paretian distribution as a statistical model for asset returns. The Fama and Mandelbrot's conjecture was supported by numerous empirical investigations in the subsequent years (see, among others, Mittnik and Rachev [21], Rachev, Ortobelli and Schwartz [22]). This paper presents and discusses conditional and unconditional portfolio selection models for returns with heavy tails.

Firstly, we examine the case of sub Gaussian α-stable distributed returns. This assumption permits a mean risk analysis pretty similar to the Markowitz-Tobin mean variance one. As a matter of fact, this model admits the same analytical form for the efficient frontier, but the parameters in the two models have a different meaning. Therefore, the most important difference is given by the way of estimating the parameters. In order to compare the performance of Gaussian and stable models we analyze an investment allocation problem. It consists of the maximization of the mean minus a measure of portfolio risk. The comparison made between the stable sub-Gaussian and the normal approach in terms of the allocation problem has indicated that the stable sub-Gaussian allocation is more risk preserving than the normal one and can give more opportunities of earning. Precisely, the stable approach, differently from the normal one, considers the component of risk due to the fat tails.

Secondly, we examine dynamic portfolio choices when returns follow an ARMA(1,1) model. Thus, in the multistage portfolio allocation problem we analyze the investor's choices considering an ARMA(1,1) model for the future scenarios of portfolio returns. Then, we compare investor's optimal allocations obtained when the residuals are either α-stable distributed or Gaussian distributed. Thus, in order to value the impact of these distributional assumptions, we propose and examine an investment allocation problem.

In Section 2 we introduce portfolio theory when returns are unconditionally stable distributed. In Section 3 we compare the stable sub-Gaussian multivariate approach with the normal multivariate one. Section 4 proposes a comparison among multi-stage conditional portfolio choice models. Finally, we briefly summarize the results.

2. The sub-Gaussian α-stable model

In this section, we analyze the problem of optimal allocation among $n+1$ assets: n of those assets are stable distributed risky assets with returns $z = [z_1, ..., z_n]'$, and the $(n+1)th$ asset is risk-free with return z_0.

Assume the vector of risky returns $z = [z_1, ..., z_n]'$ is sub-Gaussian α-stable distributed with $1 < \alpha < 2$. Then, the characteristic function of z has the following form

$$\Phi_z(t) = E(\exp(it'z)) = \exp\left(-(t'Qt)^{\frac{\alpha}{2}} + it'\mu\right), \tag{1}$$

where $Q = \left[\sigma_{ij}^2\right] = \left[\frac{q_{ij}^2}{2}\right]$ is a positive definite $(n \times n)-$matrix, $\mu = E(z)$ is the mean vector. The term q_{ij}^2 is defined by

$$\frac{q_{ij}^2}{2} = [\widetilde{z}_i, \widetilde{z}_j]_\alpha \left\|\widetilde{z}_j\right\|_\alpha^{2-\alpha}, \tag{2}$$

where $\widetilde{z}_j = z_j - \mu_j$ is the centred return, the covariation $[\widetilde{z}_i, \widetilde{z}_j]_\alpha$ between two jointly symmetric stable random variables $\widetilde{z}_i$ and $\widetilde{z}_j$ is given by

$$[\widetilde{z}_i, \widetilde{z}_j]_\alpha = \int_{S_2} s_i \left|s_j\right|^{\alpha-1} sgn(s_j)\gamma(ds),$$

in particular, $\left\|\widetilde{z}_j\right\|_\alpha = \left(\int_{S_2} \left|s_j\right|^\alpha \gamma(ds)\right)^{\frac{1}{\alpha}} = \left([\widetilde{z}_j, \widetilde{z}_j]_\alpha\right)^{\frac{1}{\alpha}}$. Here $\gamma(ds)$ is the spectral measure and it has support on the unit circle S_2.

This model can be considered as a special case of Owen-Rabinovitch's elliptical model (see Owen and Rabinovitch [19]). However, no estimate procedure of the model parameters is given in the elliptical models with infinite variance. In our approach we use (1) and (2) to provide a statistical estimator of the stable efficient frontier. To estimate the efficient frontier for returns given by (1), we need to consider an estimator for the mean vector μ and an estimator for the dispersion matrix Q. The estimator of μ is given by the vector $\widehat{\mu}$ of sample averages. Using lemma 2.7.16 in Samorodnitsky, Taqqu [27] we can write for every $p \in (1, \alpha)$

$$\frac{[\widetilde{z}_i, \widetilde{z}_j]_\alpha}{\left\|\widetilde{z}_j\right\|_\alpha^\alpha} = \frac{E\left(\widetilde{z}_i \widetilde{z}_j^{\langle p-1\rangle}\right)}{E\left(\left|\widetilde{z}_j\right|^p\right)}, \tag{3}$$

where $\widetilde{z}_j^{\langle p-1\rangle} = sgn\left(\widetilde{z}_j\right)\left|\widetilde{z}_j\right|^{p-1}$, and the scale parameter can be written $\sigma_{jj} = \left\|\widetilde{z}_j\right\|_\alpha$. Then, σ_{jj} can be approximated by the moment method suggested by Samorodnitsky, Taqqu in [27] Property 1.2.17 in the case $\beta = 0$ for every $p \in (0, \alpha)$

$$\sigma_{jj}^p = \left\|\widetilde{z}_j\right\|_\alpha^p = \frac{p\int_0^{+\infty} u^{-p-1}\sin^2 u du}{2^{p-1}\Gamma\left(1-\frac{p}{\alpha}\right)} E\left(\left|\widetilde{z}_j\right|^p\right). \tag{4}$$

Moreover the following lemma holds.

Lemma 1

For any sub-Gaussian α-stable distributed vector $\widetilde{z} = [\widetilde{z}_1, ..., \widetilde{z}_n]'$ with null mean and $1 < \alpha < 2$, it follows that $\sigma_{jj}^p = \frac{\Gamma\left(1-\frac{p}{2}\right)\sqrt{\pi}}{2^p\Gamma\left(1-\frac{p}{\alpha}\right)\Gamma\left(\frac{p+1}{2}\right)} E\left(\left|\widetilde{z}_j\right|^p\right)$ for any $p \in [0, \alpha)$ and $\frac{q_{ij}^2}{2} = \sigma_{jj}^{2-p}\frac{\Gamma\left(1-\frac{p}{2}\right)\sqrt{\pi}}{2^p\Gamma\left(1-\frac{p}{\alpha}\right)\Gamma\left(\frac{p+1}{2}\right)} E\left(\widetilde{z}_i\widetilde{z}_j^{\langle p-1\rangle}\right)$ for any $p \in [1, \alpha)$. In particular, when the vector $\widetilde{z} = [\widetilde{z}_1, ..., \widetilde{z}_n]'$ is multivariate normal distributed with null mean and variance covariance matrix $V = \left[v_{ij}^2\right]$, it follows that $v_{jj}^p = \frac{\sqrt{\pi}}{2^{\frac{p}{2}}\Gamma\left(\frac{p+1}{2}\right)} E\left(\left|\widetilde{z}_j\right|^p\right)$ for any $p \geq 0$ and $v_{ij}^2 = v_{jj}^{2-p}\frac{\sqrt{\pi}}{2^{\frac{3p-4}{2}}\Gamma\left(\frac{p+1}{2}\right)} E\left(\widetilde{z}_i\widetilde{z}_j^{\langle p-1\rangle}\right)$ for any $p \geq 1$.

Proof Suppose we have a sequence of random variables

$$X_n \in L^p(\mu) = \left\{ X | \int |X|^p \, d\mu < \infty \right\}$$

that converges in distribution to a random variable $X \in L^p(\mu)$, then the moments curves $g_n(q) = E(|X_n|^q)$ (that are analytic functions), as $n \to \infty$, converge uniformly to $g(q) = E(|X|^q)$ in the interval $[0, p]$. Consider a sequence $X_n = S_{\alpha_n}(1,0,0)$, where $\alpha_n \in (p,2)$ and $\alpha_n \nearrow 2$, then the sequence X_n converges in distribution to a Gaussian random variable X with null mean and variance equal to 2. Thus,

$$\lim_{n\to\infty} E(|X_n|^p) = \lim_{n\to\infty} \frac{2^{p-1}\Gamma\left(1-\frac{p}{\alpha_n}\right)}{p\int_0^{+\infty} u^{-p-1}\sin^2 u du} = \\ = \frac{2^{p-1}\Gamma\left(1-\frac{p}{2}\right)}{p\int_0^{+\infty} u^{-p-1}\sin^2 u du} = E(|X|^p) = \frac{2^p\Gamma\left(\frac{p+1}{2}\right)}{\sqrt{\pi}}$$

and $\int_0^{+\infty} u^{-p-1}\sin^2 u du = \frac{\Gamma\left(1-\frac{p}{2}\right)\sqrt{\pi}}{2p\Gamma\left(\frac{p+1}{2}\right)}$. Because

$$\frac{q_{ij}^2}{2} = \sigma_{jj}^2 \frac{E\left(\widetilde{z}_i \widetilde{z}_j^{\langle p-1\rangle}\right)}{E\left(|\widetilde{z}_j|^p\right)} = \sigma_{jj}^{2-p} \frac{\Gamma\left(1-\frac{p}{2}\right)\sqrt{\pi}}{2^p\Gamma\left(1-\frac{p}{\alpha}\right)\Gamma\left(\frac{p+1}{2}\right)} E\left(\widetilde{z}_i \widetilde{z}_j^{\langle p-1\rangle}\right)$$

and for any $i,j = 1,...,n$ the elements of dispersion matrix $Q = \left[\frac{q_{ij}^2}{2}\right]$ can be definded

$$\frac{q_{ij}^2}{2} = (A(\alpha,p))^{\frac{2}{p}} f(p, \widetilde{z}_i, \widetilde{z}_j) \quad \text{for every } p \in [1,\alpha) \tag{5}$$

where $A(\alpha,p) = \frac{\Gamma\left(1-\frac{p}{2}\right)\sqrt{\pi}}{2^p\Gamma\left(1-\frac{p}{\alpha}\right)\Gamma\left(\frac{p+1}{2}\right)}$, and $f(p,\widetilde{z}_i,\widetilde{z}_j) = E\left(\widetilde{z}_i \left(\widetilde{z}_j\right)^{\langle p-1\rangle}\right) \left(E(|\widetilde{z}_j|^p)\right)^{\frac{2-p}{p}}$. In addition, as $\alpha \nearrow 2$, then $Q \to \frac{V}{2}$, where V is the variance-covariance matrix and considering that $\widetilde{f}(p) = f(p,\widetilde{z}_i,\widetilde{z}_j)$ is an analytic function the thesis holds.□

The above suggests the following estimator $\widehat{Q} = \left[\frac{\widehat{q}_{ij}^2}{2}\right]$ for the entries of the unknown covariation matrix Q

$$\frac{\widehat{q}_{ij}^2}{2} = \widehat{\sigma}_{jj}^{2-p} \frac{\Gamma\left(1-\frac{p}{2}\right)\sqrt{\pi}}{2^p\Gamma\left(1-\frac{p}{\alpha}\right)\Gamma\left(\frac{p+1}{2}\right)} \frac{1}{N}\sum_{k=1}^{N} \widetilde{z}_i^{(k)} \left(\widetilde{z}_j^{(k)}\right)^{\langle p-1\rangle}, \tag{6}$$

where the σ_j is estimated as follows

$$\widehat{\sigma}_{jj}^2 = \frac{\widehat{q}_{jj}^2}{2} = \left(\frac{\Gamma\left(1-\frac{p}{2}\right)\sqrt{\pi}}{2^p\Gamma\left(1-\frac{p}{\alpha}\right)\Gamma\left(\frac{p+1}{2}\right)} \frac{1}{N}\sum_{k=1}^{N} \left|\widetilde{z}_j^{(k)}\right|^p \right)^{\frac{2}{p}}. \tag{7}$$

It is important to observe that "the best" p depends on α and on the number of observations we have. The rate of convergence of the empirical matrix $\widehat{Q} = \left[\frac{\widehat{q}_{ij}^2}{2}\right]$ to the unknown matrix Q (to be estimated), will be faster for a large

sample, if p is as small as possible, (see Rachev [20], Lamantia, Ortobelli and Rachev [13]).

Under these assumptions, the wealth $W = x'z + (1 - x'e)z_0$ associated to the portfolio x is given by

$$W = x'z + (1 - x'e)z_0 \stackrel{d}{=} S_\alpha(\sigma_W, \beta_W,\ E(W)) \quad \text{and} \quad W = z_0 \ \text{ when } \ x = 0,$$

where α is the index of stability, $\sigma_W = \sigma_{x'z} = \sqrt{x'Qx}$ is the scale (dispersion) parameter , $\beta_W = \beta_{x'z} = 0$ is the skewness parameter and $E(W) = x'E(z) + (1 - x'e)z_0$. Recall that, when $\sigma_{W_1} < \sigma_{W_2}$, $\beta_{W_1} = \beta_{W_2}$ and $E(W_1) = E(W_2)$, then W_1 second order stochastically dominates W_2 and every risk averse investor prefers W_1 to W_2 (see Ortobelli [18]). Thus, when the returns $z = [z_1, ..., z_n]'$ are jointly sub-Gaussian $\alpha-$stable distributed and unlimited short sales are allowed, every risk averse investor will choose an optimal portfolio among the portfolio solutions of the following optimization problem:

$$\begin{array}{l} \min\limits_{x} x'Qx \quad \text{subject to} \\ x'\mu + (1 - x'e)z_0 = m_W \end{array} ; \tag{8}$$

for some given mean m_W. Therefore, every optimal portfolio that maximizes a given concave utility function u, belongs to the mean-dispersion frontier

$$\sigma = \begin{cases} \frac{m - z_0}{\sqrt{(\mu - ez_0)'Q^{-1}(\mu - ez_0)}} & \text{if } \ m \geq z_0 \\ \frac{z_0 - m}{\sqrt{(\mu - ez_0)'Q^{-1}(\mu - ez_0)}} & \text{if } \ m < z_0 \end{cases} , \tag{9}$$

where $\mu = E(z)$; $m = x'\mu + (1 - x'e)z_0$; $e = [1, ..., 1]'$; and $\sigma^2 = x'Qx$. Besides, the optimal portfolio weights x satisfy the following relation:

$$x = Q^{-1}(\mu - z_0 e)\frac{m - z_0}{(\mu - ez_0)'Q^{-1}(\mu - ez_0)}. \tag{10}$$

Note that (9) and (10) have the same form as the mean-variance frontier. However, even if Q is a symmetric matrix (it is positive definite), the estimator proposed in the sub-Gaussian case (see formulas (6) and (7)) generally is not symmetric. Therefore we could obtain an inconsistent situation in which $x'z$ has stable distribution with negative squared scale parameter. However, in most of the cases $x'\widehat{Q}x > 0$ for every vector $x \in R^n$, because generally $\frac{\hat{Q} + (\hat{Q})'}{2}$ is a positive definite matrix[1].

[1] Observe that for every $x \in R^n$, we get $x'\hat{Q}x > \mathbf{0}$ if and only if $\frac{\left(\hat{Q} + (\hat{Q})'\right)}{2}$ is a positive definite matrix. Thus, we can verify that $\frac{\left(\hat{Q} + (\hat{Q})'\right)}{2}$ is positive definite in order to avoid stable portfolios $x'z$ with negative scale parameter estimators. Moreover, we observe that the symmetric matrix $\frac{\left(\hat{Q} + (\hat{Q})'\right)}{2}$ is an alternative estimator of the dispersion matrix Q whose statistical properties have to be proved.

Suppose $z = [z_1, ..., z_n]'$ is α_1-stable sub-Gaussian distributed with mean $x'\mu$ and dispersion matrix Q_{α_1}. If we approximate the vector distribution with an α_2-stable sub-Gaussian law with $1 < \alpha_2 < \alpha_1$, mean $x'\mu$ and dispersion matrix Q_{α_2}, then, $\widehat{Q}_{\alpha_2} \approx \left(\frac{A(\alpha_2,1)}{A(\alpha_1,1)}\right)^2 \widehat{Q}_{\alpha_1}$ and there are no consequences of this approximation error because the results of the portfolio selection problem (8) do not change. While if $2 \geq \alpha_2 > \alpha_1 > 1$, we cannot guarantee the same results of the portfolio choice problem. This first difference is one of the reasons for considering and studying the convergence properties of the estimator (see Rachev [20]) and the suitability of the model. Moreover, (10) exhibits the two fund separation property for both the stable and the normal case (see Ross [26]), but the matrix Q and the parameter σ have different meanings. In the normal case, Q is the variance-covariance matrix and σ is the standard deviation, while in the stable case Q is a dispersion matrix and $\sigma = \sqrt{x'Qx}$ is the scale (dispersion) parameter. According to the two-fund separation property of the sub-Gaussian α-stable approach, we can assume that the market portfolio is equal to the risky tangent portfolio under the equilibrium conditions (as in the classic mean-variance Capital Asset Pricing Model (CAPM)). Therefore, every optimal portfolio can be seen as the linear combination between the market portfolio

$$\overline{x}'z = \frac{z'Q^{-1}(\mu - z_0 e)}{e'Q^{-1}\mu - e'Q^{-1}ez_0}, \tag{11}$$

and the riskless asset return z_0. Following the same arguments as in Sharpe, Lintner, Mossin's mean-variance equilibrium model, the return of asset i is given by:

$$E(z_i) = z_0 + \beta_{i,m}(E(\overline{x}'z) - z_0), \tag{12}$$

where $\beta_{i,m} = \frac{\overline{x}'Qe^i}{\overline{x}'Q\overline{x}}$, with $e^i = [0, ..., 0, 1, 0, ..., 0]'$ the vector with 1 in the $i-th$ component and zero in all the other components.

3. A comparison between the normal multivariate distributional assumption and the stable sub-Gaussian one

In this section we examine and compare the stable sub-Gaussian assumption with the normal distributional one. Thus, we implicitly assume that returns are uniquely determined by the mean m and σ that is either the scale parameter of stable distributions or the standard deviation of normal distributions.

In a recent work Rachev, Ortobelli and Schwartz [22] compare the stable non-Gaussian assumption and the normal one by analyzing optimal allocations between a riskless return and a benchmark index. Three different indexes have been taken into consideration: CAC40, DAX30 and S&P500. Next, we extend Rachev, Ortobelli and Schwartz's comparison to the multivariate case. This comparison is formally and theoretically different from the previous one because here the benchmark index is given by the market portfolio which generally will change, if the distributional assumptions change too. Thus, as a consequence of Roll [23], [24], [25], Dybvig and Ross' [8], [9] analysis, we observe that:

a) an investor, who fits the return distributions with a joint α_1-stable sub-Gaussian distribution, will consider as inefficient the choice of another investor who fits the return distributions with a joint α_2-stable sub-Gaussian distribution with $\alpha_1 \neq \alpha_2$; and

b) the stable CAPM is still subject to some of the criticism already addressed to the classical one.

Nevertheless, it seems that the stable case better explains the empirical data. This is the main reason why we interpret and analyze the different behavior here between the investor who fits the data with joint stable sub-Gaussian distribution and the investor who fits the data with the joint normal distribution.

3.1 *An optimal allocation problem*

First, we consider the optimal allocation among 24 assets: 23 of those assets are risky assets with returns $z = [z_1, z_2, ..., z_{23}]'$ and the $24th$ is riskfree with an annual rate of 6%. We analyze the portfolio choice problems when short sales are allowed and when short sales are not allowed. In view of this comparison, we discuss and study the differences in portfolio choice problems without examining them so as to choose one of the two assumptions (Gaussian or sub-Gaussian).

In our comparison we use daily data taken from 23 international risky indexes valued in USD and quoted from January 1995 to January 1998. In the analysis proposed we first consider the maximum likelihood estimation of the stable parameters and of the Gaussian ones for every risky asset. Thus, Table I assembles the approximating parameters obtained from using Cognity System [2]. In order to compare the different stable sub-Gaussian joint distributions and the joint normal distributions for the asset returns, we assume that the vector z is sub-Gaussian α-stable distributed, with $\alpha = \alpha_k$, $k = 1, 2$, where $\alpha_1 = 1.7488$ represents the average of the indexes of stability and $\alpha_2 = 1.8856$ represents the maximum of the indexes of stability (see Table I)[3]. Moreover, when in the following tables we consider the index of stability $\alpha = 2$, we implicitly assume that the returns are jointly normal distributed. Thus, every portfolio of risky assets is stable distributed in the following way:

$$x'z \stackrel{d}{=} S_{\alpha_k}(\sigma_{x'z}, \beta_{x'z}, m_{x'z}),$$

where α_k is one of the considered indexes of stability $k = 1, 2$, $\sigma_{x'z} = (x'Q_k x)^{\frac{1}{2}}$ is the respective scale parameter, $Q_k = \left[\frac{q_{ij}^2}{2}\right]_k$ is the dispersion matrix, with $k = 1, 2$, $\beta_{x'z} = 0$ is the skewness parameter, and $m_{x'z}$ represents the mean of $x'z$. Observe that the matrix Q_k is estimated with the method defined in the previous section and thus it depends on the index of stability α_k for $k = 1, 2$. As observed previously, the rate of convergence of the empirical matrix $\widehat{Q}_k$ to the

[2] This software is developed by FinAnalytica Inc.

[3] We consider different indexes of stability, in order to value the effects of heavy - tailedness on the portfolio selection problems.

ASSETS	Gaussian Parameters		Stable Parameters			
	Mean μ	Standard Deviation σ	Index of stability α	Stable skewness β	Stable Mean μ	Stable scale parameter σ
DAX 30	0.0007	0.0113	1.8148	-0.6682	0.0005	0.0069
DAX 100	0.0007	0.0106	1.7996	-0.6389	0.0004	0.0064
CAC 40	0.0005	0.011	1.8381	-0.1852	0.0004	0.0071
FTSE all share	0.0007	0.007	1.8418	-0.5726	0.0006	0.0045
FTSE 100	0.0008	0.0078	1.8856	-0.5192	0.0007	0.0052
FTSE actuaries 350	0.0007	0.0072	1.8521	-0.5666	0.0006	0.0047
REUTERS Commodities	-0.0002	0.0072	1.7959	-0.2075	-0.0003	0.0045
NIKKEI 225 Simple Average	0.0005	0.0157	1.663	-0.0483	0.0004	0.009
NIKKEI 300 Weigh. Stock Av.	0.0006	0.0137	1.6962	0.0869	0.0006	0.0079
NIKKEI 300 Simple Stock Av.	0.0004	0.0129	1.7064	0.085	0.0004	0.0075
NIKKEI 500	0.0003	0.0128	1.7253	0.0334	0.0003	0.0076
NIKKEI 225 Stock Average	-0.0005	0.0158	1.6798	-0.0721	-0.0006	0.0091
NIKKEI 300	-0.0005	0.0138	1.6994	0.0303	-0.0005	0.008
BRENT Crude	0	0.0185	1.7423	-0.229	-0.0003	0.0112
BRENT Current Month	0	0.0186	1.7405	-0.2039	-0.0001	0.0112
CORN n. 2 Yellow cents	0.0002	0.0152	1.6869	-0.1565	0.0002	0.0083
COFFE BRAZILIAN	0.0002	0.0270	1.5876	-0.0153	0.0007	0.0144
DOW JONES FUTURES 1	-0.0001	0.0055	1.8063	-0.4641	-0.0002	0.0035
DOW JONES Commodities	-0.0001	0.0079	1.6806	-0.1389	-0.0001	0.0037
DOW JONES INDUSTRIALS	0.0013	0.0086	1.7368	-0.2886	0.0012	0.0049
FUEL OIL N. 2	-0.0001	0.0201	1.7338	-0.1961	-0.0002	0.0117
GOLDMAN SACHS Comm.	0	0.0092	1.8036	-0.2663	-0.0002	0.0058
S&P 500	0.0009	0.0083	1.7052	-0.0881	0.0010	0.0047

Table I

Maximum likelihood estimations of the Gaussian and Stable asset return parameters considering daily data from 1/3/95 to 1/30/98.

unknown matrix Q_k will be faster for a large sample, if p is as small as possible. However, studying stable simulated data we obtained very good approximations even using p not too small (we refer to Lamantia, Ortobelli and Rachev [13] for further studies on this problem). In our estimations we use $p_1 = 1.6$ (relative to $\alpha_1 = 1.7488$) and $p_2 = 1.7$ (relative to $\alpha_2 = 1.8856$).

We assume the investors wish to maximize the following utility functional:

$$U(W) = E(W) - cE\left(|W - E(W)|^q\right), \tag{13}$$

where c and q are positive real numbers, $W = \lambda z_0 + (1-\lambda)\overline{x}'z$ is the return of the portfolio, z_0 is the risk-free asset return, and $\overline{x}'z$ is the tangent portfolio return given by equation (11). With reference to the allocation problem (13), we observe that risk averse investors should choose a portfolio $W = \lambda z_0 + (1-\lambda)\overline{x}'z$ that maximizes the utility functional (13) for some real λ and some $q \in [1, \alpha)$.

We know that for $\lambda \neq 1$, portfolio return $W = \lambda z_0 + (1-\lambda)\overline{x}'z$ is distributed according to a stable law

$$S_{\alpha_k}(|1-\lambda|\,\sigma_{\overline{x}'z}, 0, \lambda z_0 + (1-\lambda)m_{\overline{x}'z}); \quad k = 1, 2$$

and $W = z_0$ when $\lambda = 1$. Now, in order to solve the asset allocation problem

$$\max_{\lambda} E(W) - cE\left(|W - E(W)|^q\right),$$

notice first that, for all $q \in [1, \alpha)$ and $1 < \alpha < 2$, we get

$$\begin{aligned} U(W) &= E(W) - cE\left(|W - E(W)|^q\right) = \\ &= \lambda z_0 + (1-\lambda)m_{\overline{x}'z} - c\left(H(\alpha, 0, q)\right)^q |1-\lambda|^q \sigma^q_{\overline{x}'z} \end{aligned}$$

where

$$\left(H(\alpha, 0, q)\right)^q = \frac{1}{A(\alpha, q)} = \frac{2^q \Gamma\left(1 - \frac{q}{\alpha}\right) \Gamma\left(\frac{q+1}{2}\right)}{\Gamma\left(1 - \frac{q}{2}\right)\sqrt{\pi}}$$

(see the above lemma and Samorodnitsky and Taqqu [27]). The above relation analyzes the stable non-Gaussian case. When the vector z admits a joint normal distribution (i.e. $\alpha = 2$), then for all $q > 0$,

$$\begin{aligned} U(W) &= E(W) - cE\left(|W - E(W)|^q\right) = \\ &= \lambda z_0 + (1-\lambda)m_{\overline{x}'z} - c\frac{2^{\frac{q}{2}}\Gamma\left(\frac{q+1}{2}\right)}{\sqrt{\pi}} |1-\lambda|^q \sigma^q_{\overline{x}'z} \end{aligned}.$$

Hence, the real optimal solution of the problem in the important case $q \in (1, \alpha)$, is given by

$$\overline{\lambda} = 1 - sgn(1-\overline{\lambda}) \left(\frac{sgn(1-\overline{\lambda})\left(m_{\overline{x}'z} - z_0\right)}{qc\sigma^q_{\overline{x}'z} V(\alpha, 0, q)} \right)^{\frac{1}{q-1}} \tag{14}$$

$$\text{and} \quad x = (1-\overline{\lambda})\overline{x}, \tag{15}$$

where $\overline{x}$ is given by (11) and

$$V(\alpha,0,q)=\begin{cases}(H(\alpha,0,q))^q & \text{in the stable case } (1<\alpha<2)\\ \frac{2^{\frac{q}{2}}\Gamma(\frac{q+1}{2})}{\sqrt{\pi}} & \text{in the normal case } (\alpha=2)\end{cases}.$$

Again, one would expect that the optimal allocation is different because the constant $V(\alpha,0,q)$ and the matrix Q are different in the stable sub-Gaussian and in the normal case.

3.2 *Stable versus normal optimal allocation: a first comparison*

We analyze the differences in optimal allocations with reference to problem (13) when the investor chooses:

1. joint normal distribution

 or

2. joint α_k stable sub-Gaussian distribution $(k=1,2)$ where $\alpha_1=1.7488$; $\alpha_2=1.8856$

as a model for the asset returns in his/her portfolio. Under these distinctive assumptions, the investors with utility functional (13) have different information about the distributional behavior of data. In particular, we examine the different market portfolio composition and the different investor's wealth allocation in the riskless asset.

First, when short sales are allowed and when short sales are not allowed, we examine optimal allocation among the riskless return and 23 index-daily returns: DAX 30, DAX 100 Performance, CAC 40, FTSE all share, FTSE 100, FTSE actuaries 350, Reuters Commodities, Nikkei 225 Simple average, Nikkei 300 weighted stock average, Nikkei 300 simple stock average, Nikkei 500, Nikkei 225 stock average, Nikkei 300, Brent Crude Physical, Brent current month, Corn No2 Yellow cents, Coffee Brazilian, Dow Jones Futures1, Dow Jones Commodities, Dow Jones Industrials, Fuel Oil No2, Goldman Sachs Commodity, S&P 500. We use the riskless return 6% p.a.. Using the estimated daily index parameters, we can compute the dispersion matrices and the approximating "market" portfolios. The dispersion matrix Q is given by either the variance-covariance matrix (in the normal case) or the matrix Q_k (in the stable cases) which depends on the index of stability α_k for $k=1,2$ ($\alpha_1=1.7488$ and $\alpha_2=1.8856$). Therefore, as shown by Tables II, III, the market portfolio weights

$$\overline{x}=\frac{Q^{-1}(\mu-z_0e)}{e'Q^{-1}\mu-e'Q^{-1}ez_0}$$

change under the different distributional assumptions. We observe that the market portfolio composition does not change excessively when we use either the asymmetric estimator (6) and (7) of matrix Q_k or the symmetric one $\frac{(\hat{Q}_k+(\hat{Q}_k)')}{2}$. However, using daily data the elements of the dispersion matrices are of order

ASSETS	Gaussian Weights (α=2)	Stable Weights for α=1.8856	Stable Weights for α=1.7488
DAX 30	1.2784	0.7182	0.3929
DAX 100	-1.1205	-0.4594	-0.0704
CAC 40	-0.2022	-0.1217	-0.0694
FTSE all share	14.0729	10.669	9.1802
FTSE 100	-1.6515	-2.5989	-3.0513
FTSE actuaries 350	-10.4307	-6.3873	-4.5968
REUTERS Commodities	-1.4679	-1.3451	-1.2833
NIKKEI 225 Simple Average	-1.8360	0.0114	0.9302
NIKKEI 300 Weigh. Stock Av.	19.791	19.4478	19.7594
NIKKEI 300 Simple Stock Av.	-11.2504	-12.0352	-12.7675
NIKKEI 500	15.5214	15.531	15.8399
NIKKEI 225 Stock Average	0.6395	-1.2088	-2.1322
NIKKEI 300	-22.4705	-21.3744	-21.2774
BRENT Crude	0.1653	0.185	0.2125
BRENT Current Month	0.046	0.026	0.0044
CORN n. 2 Yellow cents	-0.0306	0.0032	0.0295
COFFE BRAZILIAN	0.0201	0.017	0.0175
DOW JONES FUTURES 1	-0.4155	-0.2771	-0.2376
DOW JONES Commodities	-0.5327	-0.6422	-0.7123
DOW JONES INDUSTRIALS	4.0654	3.7832	3.7278
FUEL OIL N. 2	-0.2714	-0.274	-0.2806
GOLDMAN SACHS Comm.	0.2143	0.2356	0.2473
S&P 500	-3.1345	-2.9032	-2.863

Table II

Stable sub-Gaussian and Gaussian market portfolio weights when short sales are allowed

ASSETS	Gaussian Weights (α=2)	Stable Weights for α=1.8856	Stable Weights for α=1.7488
DAX 100	0.0732	0.0742	0.0746
FTSE all share	0.2496	0.2379	0.2329
NIKKEI 300 Weigh. Stock Av.	0.0485	0.0502	0.0509
DOW JONES INDUSTRIALS	0.6287	0.6377	0.6416
OTHERS	0	0	0

Table III

Stable sub-Gaussian and Gaussian market portfolio weights when no short sales are allowed

PARAMETERS	Gaussian Market Portfolio (α=2)	Stable Market Portfolio α=1.8856	Stable Market Portfolio α=1.7488
MEAN	0.0231	0.0236	0.0243
STANDARD DEVIATION	0.0418	0.0433	0.0451
DISPERSION IF α=1.8856	0.0183	0.0185	0.0191
DISPERSION IF α=1.7488	0.0101	0.0101	0.0104

Table IV

Stable sub-Gaussian and Gaussian market portfolio parameters for every mean-dispersion plane when short sales are allowed

10^{-6}. Thus the approximation in the data used could be determinant for the elements of the matrices. In particular, Table II presents the market portfolio weights when we consider all 23 asset returns and short sales are allowed. Table III gives the market portfolio weights when no short sales are allowed. Under this constraint, we value the market portfolio weights in terms of the risky portfolio compositions which maximize the extended Sharpe ratio, i.e. the market portfolio weights are the solution of the following optimization problem

$$\begin{array}{c} \max_{x} \frac{E(x'z)-z_0}{\sigma_{x'z}} \\ x'e = 1 \\ x_i \geq 0 \quad i = 1, ..., n \end{array} .$$

In this case the optimal allocation is reduced only among the four risky assets: DAX 100 Performance, FTSE all Share, Nikkei 300 weighted stock average, Dow Jones Industrials and the riskless one. As argued by Roll [23], [24], Dybvig and Ross [8], different market portfolios imply a completely different security market line analysis. Thus, the approach which takes into account short sales presents more opportunities of earning than the approach with no short sales constraint. Therefore, it dominates the other approaches. Besides, if the returns are jointly α_k stable sub-Gaussian distributed (for some determined $k = 1,2$), then the Gaussian approach is inefficient. Since, in general, efficient and inefficient portfolios can plot above and below the "real" security market line. The analysis of Tables II and III points out that the composition of the market portfolio is strictly linked to the index of stability. In fact, we see that the allocation of the market portfolio in each asset component is generally monotone with respect to the stability index. Then the intuition suggests that the stable sub-Gaussian approaches take more into consideration the component of risk because of the heavy tails. Recall that the tail behavior of every stable

PARAMETERS	Gaussian Market Portfolio (α=2)	Stable Market Portfolio α=1.8856	Stable Market Portfolio α=1.7488
MEAN	0.001072	0.001077	0.001079
STANDARD DEVIATION	0.006339	0.006376	0.006393
DISPERSION IF α=1.8856	0.002783	0.0028	0.002807
DISPERSION IF α=1.7488	0.001541	0.001549	0.001553

Table V

Stable sub-Gaussian and Gaussian market portfolio parameters for every mean-dispersion plane when no short sales are allowed

non-Gaussian distribution $X \stackrel{d}{=} S_\alpha(\sigma, \beta, \mu)$, with $1 < \alpha < 2$, is given by

$$\lim_{\lambda \to +\infty} \lambda^\alpha P(\pm X > \lambda) = C_\alpha \frac{1 \pm \beta}{2} \sigma^\alpha \tag{16}$$

where $C_\alpha = \frac{1-\alpha}{\Gamma(2-\alpha)\cos\frac{\pi\alpha}{2}}$. Therefore, the fat tails of smaller stability indexes underline the risk of the loss component of every portfolio. In particular, under the diverse distributional assumption, we distinguish the different perception of risk in the market portfolio components. This issue can be easily analyzed in the market portfolio weights with reference to the 23 returns when no short sales are allowed. In fact, Table I shows that the index of stability of FTSE all Share is greater than the other indexes of stability (of the assets DAX 100 Performance, Nikkei 300 weighted stock average, Dow Jones Industrials). Observe that in Table III the component of the FTSE all Share in the market portfolio increases with the index of stability α_k of the sub-Gaussian approach and the component of the other assets (DAX 100 Performance, Nikkei 300 weighted stock average, Dow Jones Industrials) decreases with the index of stability. Thus, the market portfolios obtained under Gaussian and sub-Gaussian distributional hypotheses consider the risks due to heavy tails differently. On the other hand the mean of market portfolios decreases with the index of stability. However, if we accept the idea that the market portfolios represent in some sense the market behavior, then according to the classic mean-risk interpretation, an optimal portfolio that has a greater mean, it has also a greater risk. This fact appears clear enough when we consider and compare the dispersion measures $\sqrt{\overline{x}_k' Q_j \overline{x}_k}$ in every mean-risk plane for every market portfolio weights $\overline{x}_k = \frac{Q_k^{-1}(\mu - z_0 e)}{e' Q_k^{-1} \mu - e' Q_k^{-1} e z_0}$, for every k and j. Observe that $\widetilde{\sigma}_{j,k} = \sqrt{\overline{x}_k' Q_j \overline{x}_k}$ is the dispersion measure of market portfolio $\overline{x}_k' z$ considering the α_j-stable Paretian approach. Therefore, for every fixed mean-risk plane (i.e. for every fixed α_j stable distributional approach) we can compare the market portfolio risk positions considering their

PARAMETERS	Gaussian Market Portfolio (α=2)	Stable Market Portfolio α=1.8856	Stable Market Portfolio α=1.7488
$VaR_{1\%}$ WITH SHORT SALES	0.0742	0.0456	0.02451
$VaR_{5\%}$ WITH SHORT SALES	0.0457	0.0212	0.00241
$VaR_{1\%}$ NO SHORT SALES	0.01366	0.0094	0.00621
$VaR_{5\%}$ NO SHORT SALES	0.00934	0.00569	0.00291

Table VI

Stable sub-Gaussian and Gaussian market portfolio Values at Risk

risk position $\widetilde{\sigma}_{j,k}$ (varying k). According to a mean-risk interpretation, we could observe that the market portfolio with a greater mean admits also a greater dispersion measure $\widetilde{\sigma}_{j,k}$, in any mean-risk plane (see tables IV and V). However, we obtain different results if we use the Value at Risk, VaR_γ, as the risk measure of the market portfolios, which is implicitly defined by the following equality:

$$VaR_\gamma(\overline{x}'z) = \sup\{y : P(\overline{x}'z \leq -y) > \gamma\}.$$

In fact, VaR measures the risk of loss which is represented by the left tail of the market portfolio distribution. Then as reported by Table VI we could observe that the $\text{VaR}_{0.01}$ and $\text{VaR}_{0.05}$ of the Gaussian market portfolio are greater than the analogous VaR_γ of the Sub-Gaussian market portfolios. In this sense the Gaussian market portfolio is riskier than the stable market portfolios because it does not consider the risk due to the heavy tails. Moreover, comparing the different VaR numbers, we can identify the market portfolio with index of stability $\alpha_1 = 1.7488$ as the least risky.

As a consequence of relation (16) it follows that every stable non-Gaussian distribution $X \stackrel{d}{=} S_\alpha(\sigma, \beta, \mu)$, with $1 < \alpha < 2$, has the property that

$$E\left(|X - E(X)|^q\right) < \infty \quad \text{for } q < \alpha \tag{17}$$

$$\text{and} \quad E\left(|X - E(X)|^q\right) = \infty \quad \text{for } q \geq \alpha \tag{18}$$

Hence, the weight of the risk measure $E\left(|X - E(X)|^q\right)$ in optimization problem (13) is generally greater for the investors who use the stable laws for asset returns when q is quite close to the index of stability α.

In Tables VII and VIII we list the optimal allocation $\overline{\lambda}$ for the normal and the stable fit. Recall that $\overline{\lambda}$ is the optimal proportion of funds invested in the risk free asset which maximizes $E(W) - cE\left(|W - E(W)|^q\right)$, where $W = \lambda z_0 + (1-\lambda)\overline{x}'z$. We have chosen $q = 1.45$ in Table VII and $q = 1.55$ in Table

Allocation $\underline{\lambda}$ in the riskless asset considering the market portfolio on 23 assets when unlimited short sales are allowed

Coefficient "*c*" of the optimization problem	Optimal allocation $\underline{\lambda}$ when α=2 (Gaussian)	Stable Optimal allocation $\underline{\lambda}$ when α=1.8856	Stable Optimal allocation $\underline{\lambda}$ when α=1.7488
c=1.8	**-0.0715**	-2.3803	*-10.2855*
c=2.2	**0.314**	-1.1642	*-6.2252*
c=3	**0.6557**	-0.0863	*-2.6268*
c=4.2	**0.837**	0.4857	*-0.7171*
c=5	**0.8893**	0.6509	*-0.1655*
c=6	**0.9262**	0.7672	*0.2227*
c=7	**0.9476**	0.8347	*0.4482*
c=10	**0.9763**	0.9252	*0.7502*
c=15	**0.9904**	0.9696	*0.8985*
c=21	**0.9954**	0.9856	*0.9520*

Allocation $\underline{\lambda}$ in the riskless asset considering the market portfolio on 23 assets when short sales are not allowed

Coefficient "*c*" of the optimization problem	Optimal allocation $\underline{\lambda}$ when α=2 (Gaussian)	Stable Optimal allocation $\underline{\lambda}$ when α=1.8856	Stable Optimal allocation $\underline{\lambda}$ when α=1.7488
c=1.4	**0.3788**	0	*0*
c=1.5	**0.4671**	0	*0*
c=1.8	**0.6446**	0	*0*
c=2.2	**0.7725**	0.3024	*0*
c=3	**0.8858**	0.6498	*0*
c=4.2	**0.9459**	0.8342	*0.4559*
c=5	**0.9633**	0.8875	*0.6306*
c=6	**0.9755**	0.925	*0.7537*
c=7	**0.9826**	0.9467	*0.8251*
c=10	**0.9921**	0.9759	*0.9208*

Table VII

This table computes the optimal allocation $\underline{\lambda}$ in the riskless return 6% annual rate (daily z_0=0.00016667) for different risk aversion coefficient *c* of the optimization problem $\max_{\lambda} E(W)-cE(|W-E(W)|^{1.45})$ where $W=\lambda z_0+(1-\lambda)\underline{x}'r$ and $\underline{x}'r$ is either the Gaussian Market portfolio (for α=2) or the sub-Gaussian market portfolio (for α=1.7488 or α=1.8856).

Allocation $\underline{\lambda}$ in the riskless asset considering the market portfolio on 23 assets when unlimited short sales are allowed

Coefficient "*c*" of the optimization problem	Optimal allocation $\underline{\lambda}$ when α=2 (Gaussian)	Stable Optimal allocation $\underline{\lambda}$ when α=1.8856	Stable Optimal allocation $\underline{\lambda}$ when α=1.7488
c=1.5	**-1.2292**	-4.5346	*-11.4475*
c=1.8	**-0.6003**	-2.973	*-7.9354*
c=2.2	**-0.1111**	-1.7584	*-5.2038*
c=3	**0.3678**	-0.5695	*-2.5298*
c=4.2	**0.6571**	0.1487	*-0.9146*
c=5	**0.7503**	0.3800	*-0.3944*
c=6	**0.8207**	0.5549	*-0.001*
c=7	**0.8646**	0.6637	*0.2437*
c=10	**0.9292**	0.8242	*0.6046*
c=15	**0.9661**	0.9159	*0.8108*

Allocation $\underline{\lambda}$ in the riskless asset considering the market portfolio on 23 assets when short sales are not allowed

Coefficient "*c*" of the optimization problem	Optimal allocation $\underline{\lambda}$ when α=2 (Gaussian)	Stable Optimal allocation $\underline{\lambda}$ when α=1.8856	Stable Optimal allocation $\underline{\lambda}$ when α=1.7488
c=1.3	**0**	0	*0*
c=1.5	**0**	0	*0*
c=1.8	**0.0859**	0	*0*
c=2.2	**0.3654**	0	*0*
c=3	**0.6389**	0.1239	*0*
c=4.2	**0.8042**	0.5248	*0*
c=5	**0.8574**	0.6539	*0.2305*
c=6	**0.8976**	0.7515	*0.4476*
c=7	**0.9226**	0.8123	*0.5826*
c=10	**0.9595**	0.9019	*0.7818*

Table VIII

This table computes the optimal allocation $\underline{\lambda}$ in the riskless return 6% annual rate (daily z_0=0.00016667) for different risk aversion coefficient c of the optimization problem $\max_{\lambda} E(W)-cE(|W-E(W)|^{1.55})$ where $W=\lambda z_0+(1-\lambda)\underline{x}'r$ and $\underline{x}'r$ is either the Gaussian Market portfolio (for α=2) or the sub-Gaussian market portfolio (for α=1.7488 or α=1.8856).

VIII, so that q is strictly less than all indexes of stability α_k, $k = 1,2$ in the data set. On the other hand, we want to value and compare the different effects of q being more distant or closer to the stability parameters α_k. For any given allocation problem, we remark in bold character and in italics respectively the greatest and the smallest allocation in the riskless asset. Both tables show that the greatest diversity is among the optimal allocations corresponding to small risk aversion coefficient c. In contrast, the very risk averse investors would take a less risky position for every distributional hypothesis and hence the allocations in the riskless asset do not change very much.

As we see from these tables, when $q = 1.45$ and $q = 1.55$ the investors who fit the data with the Gaussian approach generally assume a less risky position than the investors who fit the data with the sub-Gaussian approach. Thus, if the stable sub-Gaussian approximation presents greater performance opportunities than the Gaussian one (as observed by many empirical analyses) the "stable investors" have more opportunities of earning than the "Gaussian investors". In particular, the investors with $\alpha_1 = 1.7488$ stable sub-Gaussian approach invest less in the riskless asset than the investors who fit the data with the other approaches. However, if we consider q much closer to α_1 in the optimization problem (13), then as a consequence of (17) and (18), the investors who fit the data with α_1 stable sub-Gaussian approach assume a less risky position than the investors who fit the data with the Gaussian approach. In this case, the "stable investor" has a very risk preserving behavior because he does not prefer allocating too much wealth in the risky asset. In this sense, intuition suggests that the stable approach with lower index of stability generally is more risk preserving than with higher index of stability because the component of risk due to the fat tails of asset returns is taken into account. Therefore, the stability index plays a strategic role in the stable optimal portfolio selection. Conversely, q in the above optimization problem can be regarded as an opportune measure of the magnitude to be given to the component of risk due to the heavy tails. The importance given to q is intuitively linked to the conditions of the market in which the investor operates.

4. Portfolio selection with ARMA models

The study and the analysis of the empirical behavior of data (see, among others, Akgiray [1], Campbell [5], Fama and French [11] French, Schwert and Stambaugh [12]) have reported evidence that the conditional first and second moments of stock returns are time varying and potentially persistent especially when returns are measured over long horizons. However, the evidence observed on temporal dependence for daily stock returns is not strong (see among others Fama and French [12]) and it is still a source of controversy whether stock market price models should include long memory or not (see among others Lo [14]). In this sense the unconditional models of previous sections represent valid and indicative portfolio choice models when returns are measured over short horizons.

If financial modeling involves information on past market movements, it is not the unconditional return distribution which is of interest but the conditional distribution which is conditioned on the information contained in past return data, or in a more general information set. The class of auto-regressive moving average (ARMA) models is a natural candidate for conditioning on the past of a return series. These models, especially when the series are stationary, are widely used to predict future movements of prices, costs and returns. The reason for this popularity lies, in their simplicity and because they give a good approximation of a very wide class of stationary sequences. ARMA models have the property that the conditional distribution is homoskedastic (i.e. constant conditional volatility). In particular, if we model asset returns with an ARMA(p,q) model of auto-regressive order p and a moving average order q, returns assume the following form:

$$z_t = a_0 + \sum_{i=1}^{p} a_i z_{t-i} + \varepsilon_t + \sum_{j=1}^{q} b_j \varepsilon_{t-j},$$

where $\{\varepsilon_t\}$ is a white noise process. To specify the orders p and q, we can follow the standard Box-Jenkins identification techniques (see among others, Box and Jenkins [3], Brokwell and Davis [4]) and we inspect the sample autocorrelation functions (SACFs) and sample partial autocorrelation functions (SPACFs) of the return series. In particular, we can assume that the sequence of innovations $\{\varepsilon_t\}$ is an infinite variance process consisting of i.i.d. symmetric random variables in the domain of normal attraction of a symmetric stable distribution with index of stability $\alpha \in (0,2)$ and scale parameter $\sigma_0 > 0$. That is,

$$N^{-1/\alpha} \sum_{t=1}^{N} \varepsilon_t \xrightarrow{d} Y \quad \text{as} \quad N \longrightarrow \infty,$$

where Y has characteristic function $\Phi_Y(u) = E(e^{iuY}) = e^{-|\sigma_0 u|^\alpha}$. Under these assumptions, Mikosh, Gadrich, Kluppelberg and Adler [17] determined estimators for this process based on the sample periodogramm of z_t, and studied their asymptotic properties. Other estimators (a Gaussian-Newton type and an M-estimator) for the parameters of the ARMA process with infinite variance innovations were proposed and studied by Davis [7]. It is interesting to observe that in contrast to ARMA processes with finite variance, in the stable case we generally obtain an estimator the rate of convergence of which is considerably faster.

Next we consider a portfolio choice ARMA(1,1) model and we want to compare the impact of stable residuals and Gaussian ones. So we propose a dynamic portfolio choice among three risky indexes Dow Jones Industrial, DAX 100 and FTSE all share. We consider 5000 portfolios of these assets and we assume that portfolio returns admit the form $x'z_t = a_0 + a_1 x' z_{t-1} + \varepsilon_t + b_1 \varepsilon_{t-1}$ where we suppose the sequence of innovations $\{\varepsilon_t\}$ are either Stable or Gaussian distributed. We consider 803 daily observations of index returns from 01/04/1995 till 01/30/98 and for each portfolio we verify stationarity and we estimate the

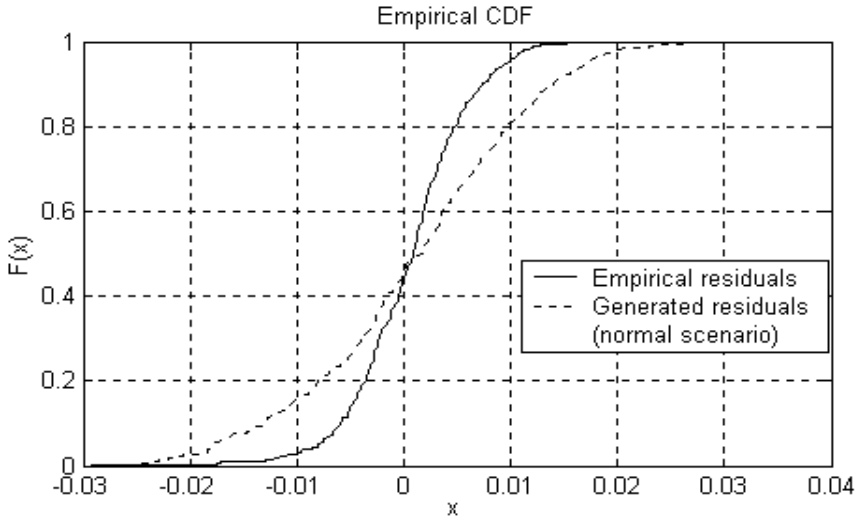


Figure 1

Kolmogorov-Smirnov test for Gaussian distributed residuals.

parameters of the model. By first comparison between the Gaussian and the stable non-Gaussian hypothesis it appears clear that stable distributions approximate much better the residuals than the Gaussian one. As a matter of fact, with the Kolmogorov-Smirnov test we can compare the empirical cumulative distribution function (cdf) $F_E(x)$ of the residuals corresponding to several portfolios with either a simulated Guassian or a simulated Stable distribution fitted in advance. By this first analysis we can generally reject at 5% confidence level the hypothesis of normality because we obtain that the probability that the empirical distribution of the residuals is Gaussian, is on average (among different portfolios) equal to $1.2\times10^{-6} \ll 5\%$. In addition we could observe that on average (among different portfolios) $k = \sup_x |F_E(x) - F(x)| = 0.1875$ where $F(x)$ is the fitted cdf of the Gaussian law. Both functions, $F_E(x)$ and $F(x)$, are shown on Figure 1.

In contrast, generally we cannot reject at 5% confidence level the hypothesis that the residuals follow a stable law because the probability, that the empirical distribution of the residuals is stable, is on average (among different portfolios) equal to 20.22%. In addition we could observe that on average (among different portfolios) $k = \sup_x |F_E(x) - F(x)| = 0.075$ when $F(x)$ is the cdf of a stable law. Both functions, $F_E(x)$ and $F(x)$, are shown on Figure 2.

In order to value the impact of different distributional approximations on

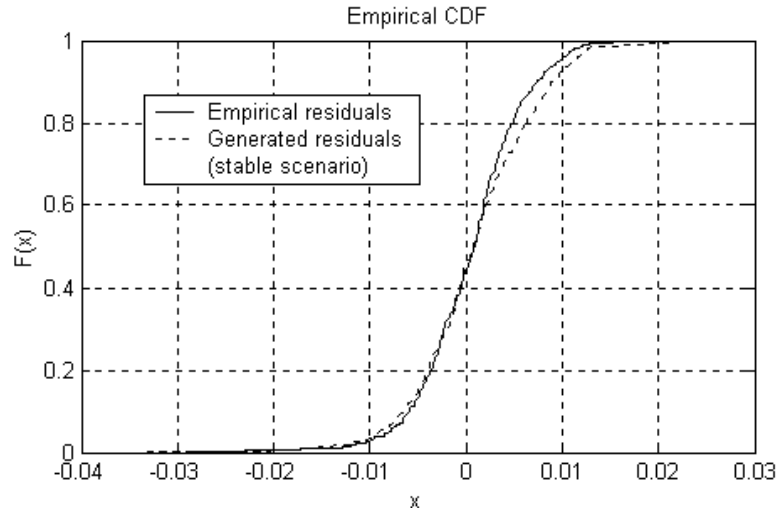


Figure 2

Kolmogorov-Smirnov test for Stable distributed residuals.

investor's portfolio choice we consider a dynamic asset allocation approach very similar to those proposed by Boender [2] and Rachev, Ortobelli and Schwartz [22]. So we generate about 2500 initial asset allocations. These allocations are then simulated into the future by using the economic scenarios, which are generated under the Gaussian and stable assumptions for the innovations of the time series models. Future economic scenarios are simulated at daily intervals. One set of scenario is generated by assuming that residuals of the variables are i.i.d. normal and another set of scenario is generated by assuming that residuals are i.i.d. stable. The horizon of interest is 10 days and two scenarios are generated for each day, so 1024 possible economic scenarios are considered for each initial portfolio. The 10-day scenario tree is repeated 10000 times. We assume that investors wish to maximize the following functional of final wealth:

$$U(W_T) = E(W_T) - cE\left(|W_T - E(W_T)|^{1.5}\right)$$

where c is a coefficient of investor's risk aversion. Therefore, the investor will choose among the initial portfolios the portfolio weight vector $x = [x_1, x_2, x_3]'$ ($x_i \geq 0$, $x_1 + x_2 + x_3 = 1$) which maximizes the utility functional $U(W_T^x)$ considering that $E(W_T^x) = \frac{1}{S}\sum_{s=1}^{S} W_{s,T}^x$ is the mean of final wealth that we obtain with portfolio x; $E\left(|W_T^x - E(W_T^x)|\right) = \frac{1}{S}\sum_{s=1}^{S}\left|W_{s,T}^x - E(W_T^x)\right|$ is the measure of risk associated to x ; $W_{s,T}^x = \Pi_{t=1}^{T}\left(1 + R_{s,t}^x\right) - 1$ is the final wealth that

Stable innovations				
Risk aversion parameter "*c*"	Expected Utility	Optimal portfolio composition Dow Jones Industrials	DAX 100	FTSE all share
0.1	0.00318	0.01643	0.96489	0.01866
0.3	-0.00286	0.01643	0.96489	0.01866
0.5	-0.006	0.06585	0.33784	0.5963
0.7	-0.00853	0.06585	0.33784	0.5963
0.8	-0.0097	0.05988	0.31699	0.62312
2.2	-0.0247	0.0813	0.21418	0.70451
Gaussian innovations				
Risk Aversion Parameter "*c*"	Expected Utility	Optimal portfolio composition Dow Jones Industrials	DAX 100	FTSE all share
0.1	0.00106	0.004192	0.520012	0.475796
0.3	-0.00295	0.04557	0.305768	0.648662
0.5	-0.0061	0.047021	0.268504	0.684475
0.7	-0.00902	0.05344	0.288484	0.658076
0.8	-0.01013	0.053643	0.239928	0.706429
2.2	-0.02575	0.053643	0.239928	0.706429

Table IX

Maximum $E(W_{10})$-$cE(|W_{10}–E(W_{10})|)$ *and portfolio composition considering return scenario generated with an ARMA*(1,1) *model under the Gaussian and stable assumptions for the innovations of the time series models.*

we obtain with portfolio x under scenario $s \in \{1, 2, ..., S\}$; $R^x_{s,t} = x_1 z_{1,s,t} + x_2 z_{2,s,t} + x_3 z_{3,s,t}$ is the return of portfolio x under scenario $s \in \{1, 2, ..., S\}$ in time period t and $z_{i,s,t}$ is the rate of return of $i-th$ asset under scenario $s \in \{1, 2, ..., S\}$ in time period t. The calculations for this empirical analysis were performed either in Matlab environment or programmed in Delphi. Table IX summarizes the results of this comparison. In particular we observe that for each risk aversion coefficient we obtain greater expected utility using stable distributed residuals. Thus it is implicitly confirmed that we have presumed a better distributional approximation. Just as in the case of the previous empirical comparison we observe that there exist substantial differences between the portfolio allocations under the different distributional approaches.

4. Conclusions

In this paper we have shown that the classical portfolio choice models can be generalized assuming stable distributions for the underlying random variables

and that the generalized models not only are theoretically justifiable and empirically testable, but they generally have better performance than the respective Gaussian models when asset returns exhibit heavy tails.

By comparing the joint normal distribution with the joint stable sub-Guassian one, it has occurred that the results received from the examined optimal allocation problems are substantially different. In particular, the stable market portfolio is generally less risky than the Gaussian market portfolio. This intuitive result is confirmed by comparison of the optimal allocations when the different distributional hypotheses are assumed. Therefore, the investors who fit the data with the stable distributions are generally more risk preserving than the investors who fit the data with the Gaussian law because stable laws take into account the component of risk due to heavy tails. Thus, we find that the tail behavior of sub-Gaussian and Gaussian approaches could imply substantial differences in the asset allocation.

These results are empirically confirmed if we compare portfolio choices obtained when considering portfolio returns that follow an ARMA(1,1) model with stable or Gaussian distributed residuals. As a matter of fact, we observed that the distribution of residuals is asymmetric and leptokurtic and the hypothesis of normality is usually rejected under statistical testing. In addition, we show that the approximation given by the stable residuals imply better performance for risk averse investors.

References

[1] V. Akgiray, Conditional heteroscedasticity in time series of stock returns: evidence and forecast, *Journal of Business,* 62, 55-80 (1989).

[2] G. C. E. Boender, Modelling and management of assets and liabilities of pension plans in the Netherlands, in *Worldwide Asset and Liability Modeling*, (W. T. Ziemba and J. M. Mulvey eds.), Cambridge University Press, Cambridge, 561-580, 1998.

[3] G. E. P. Box and G.M. Jenkins, *Time series analysis: forecasting and control,* (2nd ed.) Holden-Day, San Francisco, 1976.

[4] P. J. Brokwell, and R.A. Davis, *Time series: theory and methods,* (2nd ed.) SpringerNew York, 1991.

[5] J. Campbell, Stock returns and the term structure, *Journal of Financial Economics,* 18, 373-399, (1987).

[6] J. Chambers, S.J. Mallows and B. Stuck, A method for simulating stable random variables, *Journal of the American Statistical Association*, 71, 340-344, (1976).

[7] R. Davis, Gauss-Newton and M-estimation for ARMA processeswith infinite variance, *Stochastic Processes and Applications,* 63, 75-95, (1996).

[8] P. Dybvig and S. Ross, The analysis of performance measurement using a security market line, *Journal of Finance,* 40, 401-416, (1985a).

[9] P. Dybvig and S. Ross, Differential information and performance measurement using a security market line, *Journal of Finance,* 40, 383-399, (1985b).

[10] E. Fama, The behavior of stock market prices *Journal of Business,* 38, 34-105, (1965).

[11] E. Fama, and K. French, Permanent and temporary components of stock prices, *Journal of political Economy,* 96, 246-273, (1988).

[12] K. French, G. Schwert and R. Stambaugh, Expected stock returns and volatility, *Journal of Financial Economics,* 19, 3-29, (1987).

[13] F. Lamantia, S. Ortobelli, and S. Rachev, Value at risk with stable distributed returns, *Technical Report*, Institute of Statistics and Mathematical Economic Theory, University of Karlsruhe.(2004).

[14] A. Lo, Long term memory in stock market prices, *Econometrica,* 59, 1279-1313, (1991).

[15] B. Mandelbrot, The variation of certain speculative prices, *Journal of Business* ,26, 394-419, (1963).

[16] B. Mandelbrot, The variation of some other speculative prices, *Journal of Business,* 40, 393-413, (1967).

[17] T. Mikosh, T. Gadrich, C. Kluppelberg, and R.J. Adler, Parameter estimation for ARMA models with infinite variance innovations, *The Annals of Statistics,* 23, 305-326, (1995).

[18] S. Ortobelli, The classification of parametric choices under uncertainty: analysis of the portfolio choice problem, *Theory and Decision,* 51, 297-327, (2001).

[19] J. Owen, and R. Rabinovitch, On the class of elliptical distributions and their applications to the theory of portfolio choice, *Journal of Finance,* 38, 745-752, (1983).

[20] S. Rachev, *Probability metrics and the stability of stochastic models*, Wiley, Chichester, 1991.

[21] S. Rachev, and S. Mittnik, *Stable paretian models in finance*, Wiley, Chichester, 2000.

[22] S. Rachev, S. Ortobelli and E. Schwartz, The problem of optimal asset allocation with stable distributed returns, in Volume 238 "*Stochastic Processes and Functional Analysis a volume of recent advances in Honor of Prof. M.M. Rao*" (A.C. Krinik, R.J. Swift eds.) Marcel Dekker Inc., 295-347, 2004.

[23] R. Roll, A critique of the asset pricing theory's tests, *Journal of Financial Economics* 4, 129-176, (1977).

[24] R. Roll, Ambiguity when performance is measured by the security market line, *Journal of Finance,* 33, 1031-1069, (1978).

[25] R. Roll, Testing a portfolio for ex ante mean-variance efficiency, in *Studies in the Management Sciences* (E. Elton and M. Gruber eds.), North-Holland, Amsterdam, 1979.

[26] S. Ross, Mutual fund separation in financial theory-the separating distributions, *Journal of Economic Theory,* 17, 254-286, (1978).

[27] G. Samorodnitsky, and M. S. Taqqu, *Stable non Gaussian random processes: stochastic models with infinite variance*, Chapman and Hall, New York, 1994.

Instructions to Contributors
Journal of Concrete and Applicable Mathematics

A quartely international publication of Eudoxus Press, LLC, of TN.

Editor in Chief: George Anastassiou
Department of Mathematical Sciences
University of Memphis
Memphis, TN 38152-3240, U.S.A.

1. Manuscripts hard copies in triplicate, and in English, should be submitted to the Editor-in-Chief:

Prof.George A. Anastassiou
Department of Mathematical Sciences
The University of Memphis
Memphis,TN 38152, USA.
Tel.001. 901.678.3144
e-mail:ganastss@memphis.edu.

Authors may want to recommend an associate editor the most related to the submission to possibly handle it.
Also authors may want to submit a list of six possible referees, to be used in case we cannot find related referees by ourselves.
2. Manuscripts should be typed using any of TEX,LaTEX,AMS-TEX,or AMS-LaTEX and according to EUDOXUS PRESS, LLC. LATEX STYLE FILE. (VISIT www.msci.memphis.edu/~ganastss/jcaam / to save a copy of the style file.)They should be carefully prepared in all respects. Submitted copies should be brightly printed (not dot-matrix), double spaced, in ten point type size, on one side high quality paper 8(1/2)x11 inch. Manuscripts should have generous margins on all sides and should not exceed 24 pages.
3. Submission is a representation that the manuscript has not been published previously in this or any other similar form and is not currently under consideration for publication elsewhere. A statement transferring from the authors(or their employers,if they hold the copyright) to Eudoxus Press, LLC, will be required before the manuscript can be accepted for publication.The Editor-in-Chief will supply the necessary forms for this transfer.Such a written transfer of copyright,which previously was assumed to be implicit in the act of submitting a manuscript,is necessary under the U.S.Copyright Law in order for the publisher to carry through the dissemination of research results and reviews as widely and effective as possible.
4. The paper starts with the title of the article, author's name(s) (no titles or degrees), author's affiliation(s) and e-mail addresses. The affiliation should comprise the department, institution (usually university or company), city, state (and/or nation) and mail code.
The following items, 5 and 6, should be on page no. 1 of the paper.

5. An abstract is to be provided, preferably no longer than 150 words.
6. A list of 5 key words is to be provided directly below the abstract. Key words should express the precise content of the manuscript, as they are used for indexing purposes. The main body of the paper should begin on page no. 1, if possible.
7. All sections should be numbered with Arabic numerals (such as: 1. INTRODUCTION) .
Subsections should be identified with section and subsection numbers (such as 6.1. Second-Value Subheading).
If applicable, an independent single-number system (one for each category) should be used to label all theorems, lemmas, propositions, corrolaries, definitions, remarks, examples, etc. The label (such as Lemma 7) should be typed with paragraph indentation, followed by a period and the lemma itself.
8. Mathematical notation must be typeset. Equations should be numbered consecutively with Arabic numerals in parentheses placed flush right,and should be thusly referred to in the text [such as Eqs.(2) and (5)]. The running title must be placed at the top of even numbered pages and the first author's name, et al., must be placed at the top of the odd numbed pages.
9. Illustrations (photographs, drawings, diagrams, and charts) are to be numbered in one consecutive series of Arabic numerals. The captions for illustrations should be typed double space. All illustrations, charts, tables, etc., must be embedded in the body of the manuscript in proper, final, print position. In particular, manuscript, source, and PDF file version must be at camera ready stage for publication or they cannot be considered.
Tables are to be numbered (with Roman numerals) and referred to by number in the text. Center the title above the table, and type explanatory footnotes (indicated by superscript lowercase letters) below the table.
10. List references alphabetically at the end of the paper and number them consecutively. Each must be cited in the text by the appropriate Arabic numeral in square brackets on the baseline.

References should include (in the following order):
initials of first and middle name, last name of author(s)
title of article,
name of publication, volume number, inclusive pages, and year of publication.
Authors should follow these examples:

Journal Article

1. H.H.Gonska,Degree of simultaneous approximation of bivariate functions by Gordon operators, (journal name in italics) *J. Approx. Theory*, 62,170-191(1990).

Book

2. G.G.Lorentz, (title of book in italics) *Bernstein Polynomials* (2nd ed.), Chelsea,New York,1986.

Contribution to a Book

3. M.K.Khan, Approximation properties of beta operators,in(title of book in italics) *Progress in Approximation Theory* (P.Nevai and A.Pinkus,eds.), Academic Press, New York,1991,pp.483-495.

11. All acknowledgements (including those for a grant and financial support) should occur in one paragraph that directly precedes the References section.
12. Footnotes should be avoided. When their use is absolutely necessary, footnotes should be numbered consecutively using Arabic numerals and should be typed at the bottom of the page to which they refer. Place a line above the footnote, so that it is set

off from the text. Use the appropriate superscript numeral for citation in the text.
13. After each revision is made please again submit three hard copies of the revised manuscript, including in the final one. And after a manuscript has been accepted for publication and with all revisions incorporated, manuscripts, including the TEX/LaTex source file and the PDF file, are to be submitted to the Editor's Office on a personal-computer disk, 3.5 inch size. Label the disk with clearly written identifying information and properly ship, such as:

Your name, title of article, kind of computer used, kind of software and version number, disk format and files names of article, as well as abbreviated journal name.

Package the disk in a disk mailer or protective cardboard. Make sure contents of disks are identical with the ones of final hard copies submitted!

Note: The Editor's Office cannot accept the disk without the accompanying matching hard copies of manuscript. No e-mail final submissions are allowed! The disk submission must be used.

14. Effective 1 Nov. 2005 the journal's page charges are $8.00 per PDF file page. Upon acceptance of the paper an invoice will be sent to the contact author. The fee payment will be due one month from the invoice date. The article will proceed to publication only after the fee is paid. The charges are to be sent, by money order or certified check, in US dollars, payable to Eudoxus Press, LLC, to the address shown in the Scope and Prices section.

No galleys will be sent and the contact author will receive one(1) electronic copy of the journal issue in which the article appears.

15. This journal will consider for publication only papers that contain proofs for their listed results.

TABLE OF CONTENTS,JOURNAL OF CONCRETE AND APPLICABLE MATHEMATICS,VOL.3,NO.3,2005

ON MEROMORPHIC APPROXIMATION IN THE SPACE Lp,
A.KROT,V.PROKHOROV,……...253

PERSISTENCE AND GLOBAL STABILITY FOR PURE-DELAY TYPE NONAUTONOMOUS LOTKA-VOLTERRA DIFFERENTIAL SYSTEMS,
Y.MUROYA,……..265

A GENERAL THEORY OF STOCHASTIC ROUNDOFF ERROR ANALYSIS WITH APPLICATIONS TO DFT AND DCT,H.ZEUNER,……..283

DIFFERENTIAL PROPERTIES OF MATRIX ORTHOGONAL POLYNO-MIALS,M.CANTERO,L.MORAL,L.VELAZQUEZ,……....313

AN EULER-TYPE QUADRATURE RULE DERIVED BY USING RADON'S METHOD,C.BELINGERI,G.BRETTI,……..335

A NOTE ON QUASI POWER INCREASING SEQUENCES,H.BOR,……...347

PORTFOLIO CHOICE WITH HEAVY TAILED DISTRIBUTIONS,
S.ORTOBELLI,A.BIGLOVA,I.HUBER,B.RACHEVA,S.STOYANOV,……353

VOLUME 3,NUMBER 4 OCTOBER 2005

ISSN:1548-5390 PRINT,1559-176X ONLINE



JOURNAL

OF CONCRETE

AND APPLICABLE

MATHEMATICS

EUDOXUS PRESS,LLC

SCOPE AND PRICES OF THE JOURNAL
Journal of Concrete and Applicable Mathematics

A quartely international publication of **Eudoxus Press,LLC**

Editor in Chief: George Anastassiou
Department of Mathematical Sciences,
University of Memphis
Memphis, TN 38152, U.S.A.
ganastss@memphis.edu

The main purpose of the "Journal of Concrete and Applicable Mathematics" is to publish high quality original research articles from all subareas of Non-Pure and/or Applicable Mathematics and its many real life applications, as well connections to other areas of Mathematical Sciences, as long as they are presented in a Concrete way. It welcomes also related research survey articles and book reviews.A sample list of connected mathematical areas with this publication includes and is not restricted to: Applied Analysis, Applied Functional Analysis, Probability theory, Stochastic Processes, Approximation Theory, O.D.E, P.D.E, Wavelet, Neural Networks,Difference Equations, Summability, Fractals, Special Functions, Splines, Asymptotic Analysis, Fractional Analysis, Inequalities, Moment Theory, Numerical Functional Analysis,Tomography, Asymptotic Expansions, Fourier Analysis, Applied Harmonic Analysis, Integral Equations, Signal Analysis, Numerical Analysis, Optimization, Operations Research, Linear Programming, Fuzzyness, Mathematical Finance, Stochastic Analysis, Game Theory, Math.Physics aspects, Applied Real and Complex Analysis, Computational Number Theory, Graph Theory, Combinatorics,
Computer Science Math.related topics,combinations of the above, etc. In general any kind of Concretely presented Mathematics which is Applicable fits to the scope of this journal.
Working Concretely and in Applicable Mathematics has become a main trend in many recent years,so we can understand better and deeper and solve the important problems of our real and scientific world.
"Journal of Concrete and Applicable Mathematics" is a peer- reviewed International Quarterly Journal.
We are calling for papers for possible publication. The contributor should send three copies of the contribution to the editor in-Chief typed in TEX, LATEX double spaced. [See: Instructions to Contributors]

Journal of Concrete and Applicable Mathematics(JCAAM)
ISSN:1548-5390 PRINT, 1559-176X ONLINE.
is published in January,April,July and October of each year by
EUDOXUS PRESS,LLC,
1424 Beaver Trail Drive,Cordova,TN38016,USA,
Tel.001-901-751-3553
anastassioug@yahoo.com
http://www.EudoxusPress.com.

Annual Subscription Current Prices:For USA and Canada,Institutional:Print $250,Electronic $220,Print and Electronic $310.Individual:Print $77,Electronic $60,Print &Electronic $110.For any other part of the world add $25 more to the above prices for Print.

Single article PDF file for individual $8.Single issue in PDF form for individual $25.
The journal carries page charges $8 per page of the pdf file of an article,payable upon acceptance of the article within one month and before publication.
No credit card payments.Only certified check,money order or international check in US dollars are acceptable.
Combination orders of any two from JoCAAA,JCAAM,JAFA receive 25% discount,all three receive 30% discount.

Copyright©2004 by Eudoxus Press,LLC all rights reserved.JCAAM is printed in USA.

JCAAM is reviewed and abstracted by AMS Mathematical Reviews,MATHSCI,and Zentralblaat MATH.

It is strictly prohibited the reproduction and transmission of any part of JCAAM and in any form and by any means without the written permission of the publisher.It is only allowed to educators to Xerox articles for educational purposes.The publisher assumes no responsibility for the content of published papers.
JCAAM IS A JOURNAL OF RAPID PUBLICATION

Editorial Board
Associate Editors

Editor in -Chief:
George Anastassiou
Department of Mathematical Sciences
The University Of Memphis
Memphis,TN 38152,USA
tel.901-678-3144,fax 901-678-2480
e-mail ganastss@memphis.edu
www.msci.memphis.edu/~ganastss/jcaam
Areas:Approximation Theory,
Probability,Moments,Wavelet,
Neural Networks,Inequalities,Fuzzyness.

Associate Editors:

1) Ravi Agarwal
Florida Institute of Technology
Applied Mathematics Program
150 W.University Blvd.
Melbourne,FL 32901,USA
agarwal@fit.edu
Differential Equations,Difference
Equations,inequalities

2) Shair Ahmad
University of Texas at San Antonio
Division of Math.& Stat.
San Antonio,TX 78249-0664,USA
SAHMAD@utsa.edu
Differential Equations,Mathematical
Biology

3) Drumi D.Bainov
Medical University of Sofia
P.O.Box 45,1504 Sofia,Bulgaria
drumibainov@yahoo.com
Differential Equations,Optimal Control,
Numerical Analysis,Approximation Theory

4) Carlo Bardaro
Dipartimento di Matematica & Informatica
Universita' di Perugia
Via Vanvitelli 1
06123 Perugia,ITALY
tel.+390755855034, +390755853822,
fax +390755855024
bardaro@unipg.it ,
bardaro@dipmat.unipg.it
Functional Analysis and Approximation Th.,

19) Rupert Lasser
Institut fur Biomathematik & Biomertie,GSF
-National Research Center for environment and
health
Ingolstaedter landstr.1
D-85764 Neuherberg,Germany
lasser@gsf.de
Orthogonal Polynomials,Fourier Analysis,
Mathematical Biology

20) Alexandru Lupas
University of Sibiu
Faculty of Sciences
Department of Mathematics
Str.I.Ratiu nr.7
2400-Sibiu,Romania
lupas@ulbsibiu.ro
Classical Analysis,Inequalities,
Special Functions,Umbral Calculus,
Approximation Th.,Numerical Analysis
and Methods

21) Ram N.Mohapatra
Department of Mathematics
University of Central Florida
Orlando,FL 32816-1364
tel.407-823-5080
ramm@mail.ucf.edu
Real and Complex analysis,Approximation Th.,
Fourier Analysis, Fuzzy Sets and Systems

22) Rainer Nagel
Arbeitsbereich Funktionalanalysis
Mathematisches Institut
Auf der Morgenstelle 10
D-72076 Tuebingen
Germany
tel.49-7071-2973242
fax 49-7071-294322
rana@fa.uni-tuebingen.de
evolution equations,semigroups,spectral th.,
positivity

23) Panos M.Pardalos
Center for Appl. Optimization
University of Florida
303 Weil Hall
P.O.Box 116595
Gainesville,FL 32611-6595

Summability,Signal Analysis,Integral Equations,
Measure Th.,Real Analysis

5) Francoise Bastin
Institute of Mathematics
University of Liege
4000 Liege
BELGIUM
f.bastin@ulg.ac.be
Functional Analysis,Wavelets

6) Paul L.Butzer
RWTH Aachen
Lehrstuhl A fur Mathematik
D-52056 Aachen
Germany
tel.0049/241/80-94627 office,
0049/241/72833 home,
fax 0049/241/80-92212
Butzer@rwth-aachen.de
Approximation Th.,Sampling Th.,Signals,
Semigroups of Operators,Fourier Analysis

7) Yeol Je Cho
Department of Mathematics Education
College of Education
Gyeongsang National University
Chinju 660-701
KOREA
tel.055-751-5673 Office,
055-755-3644 home,
fax 055-751-6117
yjcho@nongae.gsnu.ac.kr
Nonlinear operator Th.,Inequalities,
Geometry of Banach Spaces

8) Sever S.Dragomir
School of Communications and Informatics
Victoria University of Technology
PO Box 14428
Melbourne City M.C
Victoria 8001,Australia
tel 61 3 9688 4437,fax 61 3 9688 4050
sever.dragomir@vu.edu.au,
sever@sci.vu.edu.au
Math.Analysis,Inequalities,Approximation Th.,
Numerical Analysis, Geometry of Banach Spaces,
Information Th. and Coding

9) A.M.Fink
Department of Mathematics
Iowa State University
Ames,IA 50011-0001,USA
tel.352-392-9011
pardalos@ufl.edu
Optimization,Operations Research

24) Svetlozar T.Rachev
Dept.of Statistics and Applied Probability Program
University of California,Santa Barbara
CA 93106-3110,USA
tel.805-893-4869
rachev@pstat.ucsb.edu
AND
Chair of Econometrics and Statistics
School of Economics and Business Engineering
University of Karlsruhe
Kollegium am Schloss,Bau II,20.12,R210
Postfach 6980,D-76128,Karlsruhe,Germany
tel.011-49-721-608-7535
rachev@lsoe.uni-karlsruhe.de
Mathematical and Empirical Finance,
Applied Probability, Statistics and Econometrics

25) Paolo Emilio Ricci
Universita' degli Studi di Roma "La Sapienza"
Dipartimento di Matematica-Istituto "G.Castelnuovo"
P.le A.Moro,2-00185 Roma,ITALY
tel.++39 0649913201,fax ++39 0644701007
riccip@uniroma1.it,Paoloemilio.Ricci@uniroma1.it
Orthogonal Polynomials and Special functions,
Numerical Analysis, Transforms,Operational Calculus,
Differential and Difference equations

26) Cecil C.Rousseau
Department of Mathematical Sciences
The University of Memphis
Memphis,TN 38152,USA
tel.901-678-2490,fax 901-678-2480
ccrousse@memphis.edu
Combinatorics,Graph Th.,
Asymptotic Approximations,
Applications to Physics

27) Tomasz Rychlik
Institute of Mathematics
Polish Academy of Sciences
Chopina 12,87100 Torun, Poland
T.Rychlik@impan.gov.pl
Mathematical Statistics,Probabilistic Inequalities

28) Bl. Sendov
Institute of Mathematics and Informatics
Bulgarian Academy of Sciences
Sofia 1090,Bulgaria

tel.515-294-8150
fink@math.iastate.edu
Inequalities,Ordinary Differential Equations

10) Sorin Gal
Department of Mathematics
University of Oradea
Str.Armatei Romane 5
3700 Oradea,Romania
galso@uoradea.ro
Approximation Th.,Fuzzyness,Complex Analysis

11) Jerome A.Goldstein
Department of Mathematical Sciences
The University of Memphis,
Memphis,TN 38152,USA
tel.901-678-2484
jgoldste@memphis.edu
Partial Differential Equations,
Semigroups of Operators

12) Heiner H.Gonska
Department of Mathematics
University of Duisburg
Duisburg,D-47048
Germany
tel.0049-203-379-3542 office
gonska@informatik.uni-duisburg.de
Approximation Th.,Computer Aided Geometric Design

13) Dmitry Khavinson
Department of Mathematical Sciences
University of Arkansas
Fayetteville,AR 72701,USA
tel.(479)575-6331,fax(479)575-8630
dmitry@uark.edu
Potential Th.,Complex Analysis,Holomorphic PDE,Approximation Th.,Function Th.

14) Virginia S.Kiryakova
Institute of Mathematics and Informatics
Bulgarian Academy of Sciences
Sofia 1090,Bulgaria
virginia@diogenes.bg
Special Functions,Integral Transforms,
Fractional Calculus

15) Hans-Bernd Knoop
Institute of Mathematics
Gerhard Mercator University
D-47048 Duisburg
Germany
tel.0049-203-379-2676
bsendov@bas.bg
Approximation Th.,Geometry of Polynomials,
Image Compression

29) Igor Shevchuk
Faculty of Mathematics and Mechanics
National Taras Shevchenko
University of Kyiv
252017 Kyiv
UKRAINE
shevchuk@univ.kiev.ua
Approximation Theory

30) H.M.Srivastava
Department of Mathematics and Statistics
University of Victoria
Victoria,British Columbia V8W 3P4
Canada
tel.250-721-7455 office,250-477-6960 home,
fax 250-721-8962
harimsri@math.uvic.ca
Real and Complex Analysis,Fractional Calculus and Appl.,
Integral Equations and Transforms,Higher Transcendental
Functions and Appl.,q-Series and q-Polynomials,
Analytic Number Th.

31) Ferenc Szidarovszky
Dept.Systems and Industrial Engineering
The University of Arizona
Engineering Building,111
PO.Box 210020
Tucson,AZ 85721-0020,USA
szidar@sie.arizona.edu
Numerical Methods,Game Th.,Dynamic Systems,
Multicriteria Decision making,
Conflict Resolution,Applications
in Economics and Natural Resources Management

32) Gancho Tachev
Dept.of Mathematics
Univ.of Architecture,Civil Eng. and Geodesy
1 Hr.Smirnenski blvd
BG-1421 Sofia,Bulgaria
Approximation Theory

33) Manfred Tasche
Department of Mathematics
University of Rostock
D-18051 Rostock
Germany
manfred.tasche@mathematik.uni-rostock.de
Approximation Th.,Wavelet,Fourier Analysis,
Numerical Methods,Signal Processing,

knoop@math.uni-duisburg.de
Approximation Theory,Interpolation

16) Jerry Koliha
Dept. of Mathematics & Statistics
University of Melbourne
VIC 3010,Melbourne
Australia
koliha@unimelb.edu.au
Inequalities,Operator Theory,
Matrix Analysis,Generalized Inverses

17) Mustafa Kulenovic
Department of Mathematics
University of Rhode Island
Kingston,RI 02881,USA
kulenm@math.uri.edu
Differential and Difference Equations

18) Gerassimos Ladas
Department of Mathematics
University of Rhode Island
Kingston,RI 02881,USA
gladas@math.uri.edu
Differential and Difference Equations
Image Processing,Harmonic Analysis

34) Chris P.Tsokos
Department of Mathematics
University of South Florida
4202 E.Fowler Ave.,PHY 114
Tampa,FL 33620-5700,USA
profcpt@math.usf.edu,profcpt@chuma1.cas.usf.edu
Stochastic Systems,Biomathematics,
Environmental Systems,Reliability Th.

35) Lutz Volkmann
Lehrstuhl II fuer Mathematik
RWTH-Aachen
Templergraben 55
D-52062 Aachen
Germany
volkm@math2.rwth-aachen.de
Complex Analysis,Combinatorics,Graph Theory

JOURNAL OF CONCRETE AND APPLICABLE MATHEMATICS,VOL.3,NO.4,389-403,
2005,COPYRIGHT 2005 EUDOXUS PRESS,LLC

Weighted Integral Inequalities in Two Dimensions

George Hanna and John Roumeliotis
School of Computer Science and Mathematics
Victoria University, Australia
Email: georgeh@csm.vu.edu.au john.roumeliotis@vu.edu.au

Abstract

Weighted (or product) double integral inequalities are developed and extended to produce weighted cubature rules. The error bounds are of first and second order and rely on the first few moments of the weight. Various properties of the weight and weight null-spaces are considered. Minimization of the bound produces coupled non-linear equations whose solution furnish optimal weighted cubature grids. These grids are evaluated for some of the more popular weight functions.

Key words: Integral inequalities, cubature, singular integration, grid generation

1. Introduction

Milovanović [3] (see also [4]), Barnett and Dragomir [1] and Hanna et al. [2] developed two dimensional integral inequalities whose error bounds were expressed in Lebesgue norms of the first partial derivatives of the integrand. In other work, Roumeliotis [7] developed and reviewed weighted one dimensional Ostrowski type inequalities with a particular emphasis of identifying optimal quadrature grids. These grids, influenced by the first few moments of the weight function, were evaluated via minimization of the Ostrowski type error bound. In this paper we combine and extend these results to develop weighted first and second order double integral inequalities. Particular attention is paid to the influence of the two dimensional weight function on the error bound and we explore this influence for different weights and weight null-spaces. Furthermore, weighted second order cubature rules are developed and we devise a method for calculating cubature grids that rely only on the first two moments of the weight. A method for calculating *a priori* cubature grids is given.

The work in this paper is presented in the following order. In Section 2, a two variable Taylor expansion is employed to develop weighted two dimensional integral inequalities. Milovanović [3] used this method to extend Ostrowski's inequality to multiple dimensions. Here we will content ourselves with two dimensions, but extend the order of the rule to two. We undertake an examination of the error bound and identify parameters that will minimize the bound. In Section 3, we present a Peano kernel method, based on analogous results in [1], to derive a second order weighted double integral inequality. Error bounds are expressed in terms of the L_1 and L_∞ norms of the first mixed partial derivative of the integrand. Particular attention is paid to minimizing this integrand for different weights and null-spaces. Finally, the results of this section are extended in Section 5 to develop a weighted cubature formula. Minimizing the error bound furnishes a set of non-linear coupled equations in the first two moments of the weight whose solution produces a cubature grid influenced by the weight function. Plots of the grid for various weights are given.

2. Taylor's Formula

In 1975, Milovanović [3] generalised the Ostrowski inequality to multiple dimensions using the multiple variable Taylor formula. As per the Ostrowski result, the inequality was expressed in terms of the first partial derivatives of the integrand. We state the two dimensional formula below.

Following [3], let $D = \{(x_1, x_2) | a_i < x_i < b_i (i = 1, 2)\}$ and let $\bar{D}$ be the closure of D.

Theorem 1. *Let $f : \mathbb{R}^2 \to \mathbb{R}$ be a differentiable function defined on $\bar{D}$ and let $\left|\frac{\partial f}{\partial t_i}\right| \le M_i$ $(M_i > 0;\ i = 1, 2)$ in D. Then, for every $X = (x_1, x_2) \in \bar{D}$,*

$$(1)\quad \left| \frac{1}{(b_1 - a_1)(b_2 - a_2)} \int_{a_1}^{b_1} \int_{a_2}^{b_2} f(t_1, t_2)\, dt_2 dt_1 - f(x_1, x_2) \right| \le M_1 (b_1 - a_1) \left(\frac{\left(x_1 - \frac{a_1 + b_1}{2}\right)^2}{(b_1 - a_1)^2} + \frac{1}{4} \right) + M_2 (b_2 - a_2) \left(\frac{\left(x_2 - \frac{a_2 + b_2}{2}\right)^2}{(b_2 - a_2)^2} + \frac{1}{4} \right).$$

The weighted version of Theorem 1 appears below.

Theorem 2. *Let $f:\mathbb{R}^2\to\mathbb{R}$ be a differentiable function defined on $\bar{D}$ and let $\left|\frac{\partial f}{\partial t_i}\right|\le M_i$ $(M_i>0;\ i=1,2)$ in D. Furthermore, let the function $X\mapsto w(X)$ be defined, integrable and $w(X)>0$ for every $X\in\bar{D}$. Then for every $X\in\bar{D}$,*

$$\left|\frac{\int_{a_1}^{b_1}\int_{a_2}^{b_2} w(t_1,t_2)f(t_1,t_2)\,dt_2dt_1}{\int_{a_1}^{b_1}\int_{a_2}^{b_2} w(t_1,t_2)\,dt_2dt_1}-f(x_1,x_2)\right| \le \frac{1}{\int_{a_1}^{b_1}\int_{a_2}^{b_2} w(t_1,t_2)\,dt_2dt_1}\left(M_1\int_{a_1}^{b_1}\int_{a_2}^{b_2} w(t_1,t_2)|x_1-t_1|\,dt_2dt_1 + M_2\int_{a_1}^{b_1}\int_{a_2}^{b_2} w(t_1,t_2)|x_2-t_2|\,dt_2dt_1\right). \tag{2}$$

Theorem 2 can be extended to higher orders and below we provide such an extension to second order.

Theorem 3. *Let $f:[a_1,b_1]\times[a_2,b_2]\to\mathbb{R}$ be such that all its partial derivatives up to order 2 exist and be continuous, i.e. $\frac{\partial^i f}{\partial t_1^j\partial t_2^k}<\infty, i=1,2; j=0,\dots,i; k=i-j$. Furthermore, let $w:(a_1,b_1)\times(a_2,b_2)\to(0,\infty)$ be integrable (i.e. $\iint w\,dA<\infty$). Then for all $(x_1,x_2)\in[a_1,b_1]\times[a_2,b_2]$ the following second order product double integral inequality holds*

$$\begin{aligned}&\left|\int_{a_1}^{b_1}\int_{a_2}^{b_2} w(t_1,t_2)f(t_1,t_2)\,dt_2dt_1 - f(x_1,x_2)\int_{a_1}^{b_1}\int_{a_2}^{b_2} w(t_1,t_2)\,dt_2dt_1\right.\\ &\left.+\frac{\partial f}{\partial t_1}(x_1,x_2)\int_{a_1}^{b_1}\int_{a_2}^{b_2} w(t_1,t_2)(x_1-t_1)\,dt_2dt_1+\frac{\partial f}{\partial t_2}(x_1,x_2)\int_{a_1}^{b_1}\int_{a_2}^{b_2} w(t_1,t_2)(x_2-t_2)\,dt_2dt_1\right|\\ &\le\frac{\|\frac{\partial^2 f}{\partial t_1^2}\|_\infty}{2}\int_{a_1}^{b_1}\int_{a_2}^{b_2} w(t_1,t_2)(x_1-t_1)^2\,dt_2dt_1+\frac{\|\frac{\partial^2 f}{\partial t_2^2}\|_\infty}{2}\int_{a_1}^{b_1}\int_{a_2}^{b_2} w(t_1,t_2)(x_2-t_2)^2\,dt_2dt_1\\ &+\left\|\frac{\partial^2 f}{\partial t_1\partial t_2}\right\|_\infty\int_{a_1}^{b_1}\int_{a_2}^{b_2} w(t_1,t_2)|x_1-t_1||x_2-t_2|\,dt_2dt_1.\end{aligned} \tag{3}$$

Proof. The two-variable Taylor formula states that

$$\begin{aligned}f(t_1,t_2)=f(x_1,x_2)&+(t_1-x_1)\frac{\partial f}{\partial t_1}(x_1,x_2)+(t_2-x_2)\frac{\partial f}{\partial t_2}(x_1,x_2)\\&+\frac{(t_1-x_1)^2}{2}\frac{\partial^2 f}{\partial t_1^2}(\xi_1,\xi_2)+(t_1-x_1)(t_2-x_2)\frac{\partial^2 f}{\partial t_1\partial t_2}(\xi_1,\xi_2)+\frac{(t_2-x_2)^2}{2}\frac{\partial^2 f}{\partial t_2^2}(\xi_1,\xi_2),\end{aligned} \tag{4}$$

where $\xi_i=t_i+\theta(x_i-t_i)$, $i=1,2$, $0<\theta<1$. Multiplying (4) by w and integrating produces the identity

$$\begin{aligned}&\int_{a_1}^{b_1}\int_{a_2}^{b_2} w(t_1,t_2)f(t_1,t_2)\,dt_2dt_1-f(x_1,x_2)\int_{a_1}^{b_1}\int_{a_2}^{b_2} w(t_1,t_2)\,dt_2dt_1\\ &+\frac{\partial f}{\partial t_1}(x_1,x_2)\int_{a_1}^{b_1}\int_{a_2}^{b_2} w(t_1,t_2)(x_1-t_1)\,dt_2dt_1\\ &+\frac{\partial f}{\partial t_2}(x_1,x_2)\int_{a_1}^{b_1}\int_{a_2}^{b_2} w(t_1,t_2)(x_2-t_2)\,dt_2dt_1\\ &=\int_{a_1}^{b_1}\int_{a_2}^{b_2} w(t_1,t_2)\frac{(t_1-x_1)^2}{2}\frac{\partial^2 f}{\partial t_1^2}(\xi_1,\xi_2)\,dt_2dt_1\\ &+\int_{a_1}^{b_1}\int_{a_2}^{b_2} w(t_1,t_2)(t_1-x_1)(t_2-x_2)\frac{\partial^2 f}{\partial t_1\partial t_2}(\xi_1,\xi_2)\,dt_2dt_1\\ &+\int_{a_1}^{b_1}\int_{a_2}^{b_2} w(t_1,t_2)\frac{(t_2-x_2)^2}{2}\frac{\partial^2 f}{\partial t_2^2}(\xi_1,\xi_2),\,dt_2dt_1.\end{aligned} \tag{5}$$

Taking the modulus of both sides of (5), applying the triangle inequality and then Hölder's inequality on the right hand side gives (3). □

Corollary 4. *Let the conditions for f be as in Theorem 3. Then the following double integral inequality holds*

$$\begin{aligned}(6)\quad &\left|\frac{1}{(b_1-a_1)(b_2-a_2)}\int_{a_1}^{b_1}\int_{a_2}^{b_2} f(t_1,t_2)\,dt_2dt_1 - f(x_1,x_2) + \frac{\partial f}{\partial t_1}(x_1,x_2)\left(x_1-\frac{a_1+b_1}{2}\right)\right.\\ &\left. + \frac{\partial f}{\partial t_2}(x_1,x_2)\left(x_1-\frac{a_1+b_1}{2}\right)\right| \le \left\|\frac{\partial^2 f}{\partial t_1^2}\right\|_\infty \frac{(b_1-a_1)^2}{2}\left(\frac{\left(x_1-\frac{a_1+b_1}{2}\right)^2}{(b_1-a_1)^2}+\frac{1}{12}\right)\\ &+\left\|\frac{\partial^2 f}{\partial t_1\partial t_2}\right\|_\infty (b_1-a_1)(b_2-a_2)\left(\frac{\left(x_1-\frac{a_1+b_1}{2}\right)^2}{(b_1-a_1)^2}+\frac{1}{4}\right)\left(\frac{\left(x_2-\frac{a_2+b_2}{2}\right)^2}{(b_2-a_2)^2}+\frac{1}{4}\right)\\ &+\left\|\frac{\partial^2 f}{\partial t_2^2}\right\|_\infty \frac{(b_2-a_2)^2}{2}\left(\frac{\left(x_2-\frac{a_2+b_2}{2}\right)^2}{(b_2-a_2)^2}+\frac{1}{12}\right).\end{aligned}$$

Proof. Substituting $w(t_1,t_2)=1$ into (3) and simplifying produces the desired result. □

The point (x_1,x_2), the sample point of the integration rule, is free to be chosen. Often, such points are chosen to simplify the rule. For example, in (3) if we choose the *weight mean*

$$x_i = \frac{\int_{a_1}^{b_1}\int_{a_2}^{b_2} t_i w(t_1,t_2)\,dt_2dt_1}{\int_{a_1}^{b_1}\int_{a_2}^{b_2} w(t_1,t_2)\,dt_2dt_1}, \quad i=1,2$$

then the partial derivative terms vanish. Fortuitously, in this case, this point also minimizes the bound. In the following sub-section, and indeed this paper, we will not be concerned with simplifying the integration rule, but instead attempt to determine such parameters (for eg. x_1 and x_2) in order for the error bound to be minimized.

2.1. **Minimizing the upper bound.**

Corollary 5. *The bound in equation (2) is minimized at the median point (x_1,x_2) satisfying*

$$(7)\qquad \int_{a_1}^{x_1}\int_{a_2}^{b_2} w(t_1,t_2)\,dt_2dt_1 = \int_{x_1}^{b_1}\int_{a_2}^{b_2} w(t_1,t_2)\,dt_2dt_1$$

and

$$(8)\qquad \int_{a_2}^{x_2}\int_{a_1}^{b_1} w(t_1,t_2)\,dt_1dt_2 = \int_{x_2}^{b_2}\int_{a_1}^{b_1} w(t_1,t_2)\,dt_1dt_2.$$

Proof. It is a simple matter to show that

$$I(x_1,x_2) = M_1\int_{a_1}^{b_1}\int_{a_2}^{b_2} w(t_1,t_2)|x_1-t_1|\,dt_2dt_1 + M_2\int_{a_1}^{b_1}\int_{a_2}^{b_2} w(t_1,t_2)|x_2-t_2|\,dt_2dt_1$$

is a convex function. Hence the upper bound in (2) is minimized at the stationary point of I. Evaluating the first partial derivatives of I produces equations (7) and (8). □

That is, the minimum point is the *median* of the weight in each direction. This is consistent with first order rules reported in [7].

Minimization of the second order bound in Theorem 3 is not as simple. It is quite difficult to identify a minimum point for the upper bound of (3). This bound is comprised of three components; the first and last are minimized at the mean (in each direction)

$$(9)\qquad x_1 = \frac{\int_{a_1}^{b_1}\int_{a_2}^{b_2} t_1 w(t_1,t_2)\,dt_2dt_1}{\int_{a_1}^{b_1}\int_{a_2}^{b_2} w(t_1,t_2)\,dt_2dt_1}, \qquad x_2 = \frac{\int_{a_1}^{b_1}\int_{a_2}^{b_2} t_2 w(t_1,t_2)\,dt_2dt_1}{\int_{a_1}^{b_1}\int_{a_2}^{b_2} w(t_1,t_2)\,dt_2dt_1},$$

while the second is minimized at the root of a median-type expression

$$(10)\qquad \int_{a_1}^{x_1}\int_{a_2}^{b_2} |x_2-t_2| w(t_1,t_2)\,dt_2dt_1 = \int_{x_1}^{b_1}\int_{a_2}^{b_2} |x_2-t_2| w(t_1,t_2)\,dt_2dt_1$$

and

$$(11)\qquad \int_{a_2}^{x_2}\int_{a_1}^{b_1} |x_1-t_1| w(t_1,t_2)\,dt_1dt_2 = \int_{x_2}^{b_2}\int_{a_1}^{b_1} |x_1-t_1| w(t_1,t_2)\,dt_1dt_2.$$

Of course, for weights in which the solutions of (9) are identical to those of (10) and (11) then identification of the minimum point presents little challenge. For example if w is a product weight and symmetric about the midpoint $(\frac{a_1+b_1}{2},\frac{a_2+b_2}{2})$ then the minimum point is the midpoint. That is, if $w(t_1,t_2)=w_1(t_1)w_2(t_2)$

and $w_i((a+b)/2-t) = w_i((a+b)/2+t)$ $(i=1,2)$, then it can be shown that the solution of (9)–(11) is the mid-point. This is the case when $w=1$ and Corollary 4 shows that the upper bound is minimized at $x_i = (a_i+b_i)/2$, $i=1,2$.

The major difficulty with (3) is that the upper bound is comprised of a linear combination of three terms involving norms of the partial derivative of the integrand. Hence it would be near impossible to find a global minimum that depends only on the weight and not f. To obtain a global minimum for a general second order rule will require either simplification of (3) or the derivation of another expression for the bound. The first point is dealt with in the corollary below, while the second is taken up in the next section.

Corollary 6. *Let f and w be as given in Theorem 3. Then for all $(x_1,x_2) \in [a_1,b_1]\times[a_2,b_2]$ the following second order product double integral inequality holds*

$$\begin{aligned}
&\left| \int_{a_1}^{b_1}\int_{a_2}^{b_2} w(t_1,t_2)f(t_1,t_2)\,dt_2dt_1 - f(x_1,x_2)\int_{a_1}^{b_1}\int_{a_2}^{b_2} w(t_1,t_2)\,dt_2dt_1 \right.\\
&\left. + \frac{\partial f}{\partial t_1}(x_1,x_2)\int_{a_1}^{b_1}\int_{a_2}^{b_2} w(t_1,t_2)(x_1-t_1)\,dt_2dt_1 + \frac{\partial f}{\partial t_2}(x_1,x_2)\int_{a_1}^{b_1}\int_{a_2}^{b_2} w(t_1,t_2)(x_2-t_2)\,dt_2dt_1 \right|\\
&\le \left\|\frac{\partial^2 f}{\partial t_1^2}\right\|_\infty \frac{\|w\|_1}{2}\left[\left|x_1-\frac{a_1+b_1}{2}\right|+\frac{b_1-a_1}{2}\right]^2 + \left\|\frac{\partial^2 f}{\partial t_2^2}\right\|_\infty \frac{\|w\|_1}{2}\left[\left|x_2-\frac{a_2+b_2}{2}\right|+\frac{b_2-a_2}{2}\right]^2\\
&\quad + \left\|\frac{\partial^2 f}{\partial t_1\partial t_2}\right\|_\infty \|w\|_1\left[\left|x_1-\frac{a_1+b_1}{2}\right|+\frac{b_1-a_1}{2}\right]\left[\left|x_2-\frac{a_2+b_2}{2}\right|+\frac{b_2-a_2}{2}\right], \qquad (12)
\end{aligned}$$

where $\|w\|_1 = \int_{a_1}^{b_1}\int_{a_2}^{b_2} w(t_1,t_2)\,dt_2dt_1$ is the zero-th moment of the weight.

Proof. The proof involves taking an upper bound of (3) using Hölder's inequality. Thus, consider

$$\begin{aligned}
\int_{a_1}^{b_1}\int_{a_2}^{b_2} w(t_1,t_2)(x_1-t_1)^2\,dt_2dt_1 &\le \sup_{t_1\in[a_1,b_1]}(x_1-t_1)^2\int_{a_1}^{b_1}\int_{a_2}^{b_2} w(t_1,t_2)\,dt_2dt_1\\
&= \max\{(x_1-a_1)^2,(x_1-b_1)^2\}\|w\|_1\\
&= \left[\left|x_1-\frac{a_1+b_1}{2}\right|+\frac{b_1-a_1}{2}\right]^2\|w\|_1. \qquad (13)
\end{aligned}$$

Similarly

$$\int_{a_1}^{b_1}\int_{a_2}^{b_2} w(t_1,t_2)(x_2-t_2)^2\,dt_2dt_1 \le \left[\left|x_2-\frac{a_2+b_2}{2}\right|+\frac{b_2-a_2}{2}\right]^2\|w\|_1. \qquad (14)$$

Finally,

$$\begin{aligned}
&\int_{a_1}^{b_1}\int_{a_2}^{b_2} w(t_1,t_2)|x_1-t_1||x_2-t_2|\,dt_2dt_1\\
&\quad\le \sup_{(t_1,t_2)\in[a_1,b_1]\times[a_2,b_2]}|x_1-t_1||x_2-t_2|\|w\|_1\\
&\quad= \max\{x_1-a_1,b_1-x_1\}\max\{x_2-a_2,b_2-x_2\}\|w\|_1\\
&\quad= \left[\left|x_1-\frac{a_1+b_1}{2}\right|+\frac{b_1-a_1}{2}\right]\left[\left|x_2-\frac{a_2+b_2}{2}\right|+\frac{b_2-a_2}{2}\right]\|w\|_1. \qquad (15)
\end{aligned}$$

Making use of (13), (14) and (15) gives (3). □

It is clear that the bound in (12) is minimized at the mid-point of the rectangular region. Unfortunately, the weight does not influence this minimum point.

Taylor's theorem is a popular vehicle for developing cubature and higher dimension rules. Stroud [8] uses Taylor's expansion to develop cubature rules and recently Qi [5], used this technique to derive weighted Iyengar-type multiple integrals. The drawback is in the size of the error bound. For two dimensions, an n-th order rule has a Taylor remainder of $n+1$ terms. Minimizing any rule with order greater than one would be extremely difficult. Thus, in the next section, we turn to the Peano kernel and use the results of [1, 2] to derive a second order weighted double integral inequality that contains only one term in the upper bound.

3. Main Results

Lemma 7. *Let $f:[a_1,b_1]\times[a_2,b_2]\to\mathbb{R}$ be bounded and integrable and whose first partial derivatives exist and are also bounded and integrable. Furthermore, let $w:(a_1,b_1)\times(a_2,b_2)\to(0,\infty)$ be integrable. The following*

identity holds

$$I=\int_{a_1}^{b_1}\int_{a_2}^{b_2}\left[f\left(x_1,x_2\right)-f\left(x_1,t_2\right)-f\left(t_1,x_2\right)+f\left(t_1,t_2\right)\right]w\left(t_1,t_2\right)\,dt_2dt_1 \tag{16}$$
$$=\int_{a_1}^{b_1}\int_{a_2}^{b_2}P(t_1,t_2)\frac{\partial^2 f}{\partial t_1\partial t_2}\,dt_2dt_1$$

where $x_1\in[a_1,b_1]$, $x_2\in[a_2,b_2]$ *and*

$$P(t_1,t_2)=\begin{cases}\displaystyle\int_{a_2}^{t_2}p\left(t_1,u_2\right)du_2, & a_2\le t_2\le x_2,\\ \displaystyle\int_{b_2}^{t_2}p\left(t_1,u_2\right)du_2, & x_2<t_2\le b_2,\end{cases} \tag{17}$$

$$p(t_1,t_2)=\begin{cases}\displaystyle\int_{a_1}^{t_1}w\left(u_1,t_2\right)du_1, & a_1\le t_1\le x_1,\\ \displaystyle\int_{b_1}^{t_1}w\left(u_1,t_2\right)du_1, & x_1<t_1\le b_1.\end{cases} \tag{18}$$

Proof. To begin, let $I=\int_{a_1}^{b_1}I_2\,dt_1$ and consider I_2 where

$$\begin{aligned}I_2&=\int_{a_2}^{b_2}P(t_1,t_2)\frac{\partial^2 f\left(t_1,t_2\right)}{\partial t_1\partial t_2}\,dt_2\\&=\int_{a_2}^{x_2}P(t_1,t_2)\frac{\partial^2 f\left(t_1,t_2\right)}{\partial t_1\partial t_2}\,dt_2+\int_{x_2}^{b_2}P(t_1,t_2)\frac{\partial^2 f\left(t_1,t_2\right)}{\partial t_1\partial t_2}\,dt_2\\&=\int_{a_2}^{x_2}\left(\int_{a_2}^{t_2}p(t_1,u_2)du_2\right)\frac{\partial^2 f\left(t_1,t_2\right)}{\partial t_1\partial t_2}\,dt_2+\int_{x_2}^{b_2}\left(\int_{b_2}^{t_2}p(t_1,u_2)du_2\right)\frac{\partial^2 f\left(t_1,t_2\right)}{\partial t_1\partial t_2}\,dt_2\\&=I_{21}+I_{22}.\end{aligned}$$

Using integration by parts, we find that

$$\begin{aligned}I_{21}&=\int_{a_2}^{t_2}p(t_1,u_2)du_2\left.\frac{\partial f\left(t_1,t_2\right)}{\partial t_1}\right|_{a_2}^{x_2}-\int_{a_2}^{x_2}\frac{\partial f\left(t_1,t_2\right)}{\partial t_1}p(t_1,t_2)\,dt_2\\&=\int_{a_2}^{x_2}p(t_1,u_2)du_2\frac{\partial f\left(t_1,x_2\right)}{\partial t_1}-\int_{a_2}^{x_2}\frac{\partial f\left(t_1,t_2\right)}{\partial t_1}p(t_1,t_2)\,dt_2\\&=\int_{a_2}^{x_2}p(t_1,t_2)\left(\frac{\partial f\left(t_1,x_2\right)}{\partial t_1}-\frac{\partial f\left(t_1,t_2\right)}{\partial t_1}\right)dt_2.\end{aligned}$$

Similarly

$$I_{22}=\int_{x_2}^{b_2}p(t_1,t_2)\left(\frac{\partial f\left(t_1,x_2\right)}{\partial t_1}-\frac{\partial f\left(t_1,t_2\right)}{\partial t_1}\right)dt_2.$$

Thus I_2 becomes

$$I_2=\int_{a_2}^{b_2}p(t_1,t_2)\left(\frac{\partial f\left(t_1,x_2\right)}{\partial t_1}-\frac{\partial f\left(t_1,t_2\right)}{\partial t_1}\right)dt_2$$

and substituting into I gives

$$\begin{aligned}I=\int_{a_1}^{b_1}\int_{a_2}^{b_2}P(t_1,t_2)\frac{\partial^2 f}{\partial t_1\partial t_2}\,dt_2dt_1&=\int_{a_1}^{b_1}\int_{a_2}^{b_2}p(t_1,t_2)\left(\frac{\partial f\left(t_1,x_2\right)}{\partial t_1}-\frac{\partial f\left(t_1,t_2\right)}{\partial t_1}\right)dt_2dt_1\\&=\int_{a_2}^{b_2}\int_{a_1}^{b_1}p(t_1,t_2)\left(\frac{\partial f\left(t_1,x_2\right)}{\partial t_1}-\frac{\partial f\left(t_1,t_2\right)}{\partial t_1}\right)dt_1dt_2\\&=\int_{a_2}^{b_2}I_3\,dt_2,\end{aligned} \tag{19}$$

where

$$I_3=\int_{a_1}^{b_1}p(t_1,t_2)\left(\frac{\partial f\left(t_1,x_2\right)}{\partial t_1}-\frac{\partial f\left(t_1,t_2\right)}{\partial t_1}\right)dt_1.$$

Applying the same treatment to I_3 as for I_2 gives

$$I_3=\int_{a_1}^{b_1}w(t_1,t_2)[f(x_1,x_2)-f(t_1,x_2)-f(x_1,t_2)+f(t_1,t_2)]\,dt_1.$$

Substituting I_3 into (19) we find that the identity (16) is thus proved. □

The upper bound of the integration rule will depend on P. Below, we detail some properties of P that will be subsequently used in analysis of the bound.

Lemma 8. *The kernel $P : [a_1, b_1] \times [a_2, b_2] \to \mathbb{R}$ as defined in Lemma 7 has the following properties:*

(1) *P vanishes on the boundary of the rectangle $[a_1, b_1] \times [a_2, b_2]$,*
(2) *$P(t_1, \cdot) : (a_2, b_2) \to \mathbb{R}$ is monotonic increasing for all $t_1 \in (a_1, x_1)$,*
(3) *$P(t_1, \cdot) : (a_2, b_2) \to \mathbb{R}$ is monotonic decreasing for all $t_1 \in (x_1, b_1)$,*
(4) *P is positive on $(a_1, x_1) \times (a_2, x_2)$ and $(x_1, b_1) \times (x_2, b_2)$,*
(5) *P is negative on $(a_1, x_1) \times (x_2, b_2)$ and $(x_1, b_1) \times (a_2, x_2)$,*

for all $(x_1, x_2) \in (a_1, b_1) \times (a_2, b_2)$.

Proof. These properties are quite simple to prove via inspection of the first partial derivatives of P. □

In Figure 1, we plot the surface and contours of (17) for two different weights. The plots exhibit the properties discussed in Lemma 8. It is obvious that the kernel achieves its maximum deviation on of its branches at the discontinuous point (x_1, x_2).

In the following theorem we state the main result by employing the identity in Lemma 7 to produce second order weighted double integral inequalities. In contrast with the inequalities of the previous section, the upper bound here is comprised of just one term.

Theorem 9. *Let the conditions of Lemma 7 hold. The following double integral inequalities involving the usual Lebesgue norms of the first mixed partial derivative of f hold,*

$$|I| \leq \left\| \frac{\partial^2 f}{\partial t_1 \partial t_2} \right\|_\infty \int_{a_1}^{b_1} \int_{a_2}^{b_2} |x_1 - t_1| \, |x_2 - t_2| \, w(t_1, t_2) \, dt_1 dt_2, \tag{20}$$

if $\frac{\partial^2 f}{\partial t_1 \partial t_2} \in L_\infty [a_1, b_1] \times [a_2, b_2]$ and

$$\begin{aligned} |I| \leq \left\| \frac{\partial^2 f}{\partial t_1 \partial t_2} \right\|_1 \max\Bigg\{ & \int_{a_1}^{x_1} \int_{a_2}^{x_2} w(t_1, t_2) \, dt_2 dt_1, \int_{a_1}^{x_1} \int_{x_2}^{b_2} w(t_1, t_2) \, dt_2 dt_1, \\ & \int_{x_1}^{b_1} \int_{a_2}^{x_2} w(t_1, t_2) \, dt_2 dt_1, \int_{x_1}^{b_1} \int_{x_2}^{b_2} w(t_1, t_2) \, dt_2 dt_1 \Bigg\} \end{aligned} \tag{21}$$

if $\frac{\partial^2 f}{\partial t_1 \partial t_2} \in L_1 [a_1, b_1] \times [a_2, b_2]$, where I is defined in equation (16).

Proof. To prove (20) we begin with Hölder's inequality and then simplify using Lemma 8

$$\begin{aligned} |I| &= \left| \int_{a_1}^{b_1} \int_{a_2}^{b_2} P(t_1, t_2) \frac{\partial^2 f(t_1, t_2)}{\partial t_1 \partial t_2} dt_1 dt_2 \right| \\ &\leq \left\| \frac{\partial^2 f}{\partial t_1 \partial t_2} \right\|_\infty \int_{a_1}^{b_1} \int_{a_2}^{b_2} |P(t_1, t_2)| \, dt_2 dt_1 \\ &= \left\| \frac{\partial^2 f}{\partial t_1 \partial t_2} \right\|_\infty \Bigg(\int_{a_1}^{x_1} \int_{a_2}^{x_2} P(t_1, t_2) \, dt_2 dt_1 - \int_{a_1}^{x_1} \int_{x_2}^{b_2} P(t_1, t_2) \, dt_2 dt_1 \\ &\qquad - \int_{x_1}^{b_1} \int_{a_2}^{x_2} P(t_1, t_2) \, dt_2 dt_1 + \int_{x_1}^{b_1} \int_{x_2}^{b_2} P(t_1, t_2) \, dt_2 dt_1 \Bigg). \end{aligned} \tag{22}$$

Now each of the terms in (22) can be evaluated via partial integration and simplified using Lemma 7 and equations (17) and (18). For the first term

$$\begin{aligned} \int_{a_1}^{x_1} \int_{a_2}^{x_2} P(t_1, t_2) \, dt_2 dt_1 &= \int_{a_1}^{x_1} \Big\{ (t_2 - x_2) P \big|_{a_2}^{x_2} - \int_{a_2}^{x_2} (t_2 - x_2) p \, dt_2 \Big\} \, dt_1 \\ &= - \int_{a_2}^{x_2} \int_{a_1}^{x_1} (t_2 - x_2) p \, dt_1 dt_2 \\ &= - \int_{a_2}^{x_2} (t_2 - x_2) \Big\{ (t_1 - x_1) p \big|_{a_1}^{x_1} - \int_{a_1}^{x_1} (t_1 - x_1) w \, dt_1 \Big\} \, dt_2 \\ &= \int_{a_2}^{x_2} \int_{a_1}^{x_1} (x_2 - t_2)(x_1 - t_1) w(t_1, t_2) \, dt_1 dt_2. \end{aligned} \tag{23}$$

Employing the same procedure for the other terms we find

$$\int_{a_1}^{x_1} \int_{x_2}^{b_2} P(t_1, t_2) \, dt_2 dt_1 = \int_{a_1}^{x_1} \int_{x_2}^{b_2} (x_2 - t_2)(x_1 - t_1) w(t_1, t_2) \, dt_1 dt_2, \tag{24}$$

$$\int_{x_1}^{b_1}\int_{a_2}^{x_2} P(t_1,t_2)\,dt_2dt_1 = \int_{x_1}^{b_1}\int_{a_2}^{x_2}(x_2-t_2)(x_1-t_1)w(t_1,t_2)\,dt_1dt_2, \tag{25}$$

$$\int_{x_1}^{b_1}\int_{x_2}^{b_2} P(t_1,t_2)\,dt_2dt_1 = \int_{x_1}^{b_1}\int_{x_2}^{b_2}(x_2-t_2)(x_1-t_1)w(t_1,t_2)\,dt_1dt_2. \tag{26}$$

Substituting (23)–(26) into (22) gives (20). To prove (21) we again begin with Hölder's inequality

$$\begin{aligned}|I| &= \left|\int_{a_1}^{b_1}\int_{a_2}^{b_2} P(t_1,t_2)\frac{\partial^2 f\,(t_1,t_2)}{\partial t_1\partial t_2}dt_1dt_2\right| \\ &\le \left\|\frac{\partial^2 f}{\partial t_1\partial t_2}\right\|_1 \sup_{(t_1,t_2)\in[a_1,b_1]\times[a_2,b_2]} |P(t_1,t_2)| \\ &= \left\|\frac{\partial^2 f}{\partial t_1\partial t_2}\right\|_1 \max\Bigg\{\int_{a_1}^{x_1}\int_{a_2}^{x_2} w(t_1,t_2)\,dt_2dt_1, \int_{a_1}^{x_1}\int_{x_2}^{b_2} w(t_1,t_2)\,dt_2dt_1, \\ &\qquad \int_{x_1}^{b_1}\int_{a_2}^{x_2} w(t_1,t_2)\,dt_2dt_1, \int_{x_1}^{b_1}\int_{x_2}^{b_2} w(t_1,t_2)\,dt_2dt_1\Bigg\}. \end{aligned} \tag{27}$$

The last line being computed by appealing to the properties of P as listed in Lemma 8. Thus the theorem is proved. □

If the first moments of the weight w are known, as well as the *one dimensional* integrals

$$\int_{a_1}^{b_1} f(t_1,x_2)\left(\int_{a_2}^{b_2} w(t_1,t_2)\,dt_2\right)dt_1 \quad\text{and}\quad \int_{a_2}^{b_2} f(x_1,t_2)\left(\int_{a_1}^{b_1} w(t_1,t_2)\,dt_1\right)dt_2 \tag{28}$$

then (20) can form the basis of a cubature formula for the evaluation of the weighted double integral $\iint_D f(t_1,t_2)w(t_1,t_2)\,dA$ over a rectangular region D. A major drawback is that in most cases the integrals (28) are unknown. These can be eliminated using the one-dimensional weighted results in [6]. Roumeliotis *et al.* [6] showed that for mappings f with bounded second derivative that

$$\left|\int_a^b w(t)f(t)\,dt - f(x)\int_a^b w(t)\,dt + f'(x)\int_a^b (x-t)w(t)\,dt\right| \le \frac{\|f''\|_\infty}{2}\int_a^b (x-t)^2 w(t)\,dt, \tag{29}$$

where $x\in(a,b)$ and w is a weight function. Thus making use of (29), the following inequalities hold

$$\begin{aligned}&\left|\int_{a_1}^{b_1} f(t_1,x_2)\left(\int_{a_2}^{b_2} w(t_1,t_2)\,dt_2\right)dt_1 - f(x_1,x_2)\int_{a_1}^{b_1}\int_{a_2}^{b_2} w(t_1,t_2)\,dt_2\,dt_1\right. \\ &\qquad \left. + \frac{\partial f}{\partial t_1}(x_1,x_2)\int_{a_1}^{b_1}\int_{a_2}^{b_2}(x_1-t_1)w(t_1,t_2)\,dt_2\,dt_1\right| \le \left\|\frac{\partial^2 f}{\partial t_1^2}\right\|_\infty \int_{a_1}^{b_1}\int_{a_2}^{b_2} w(t_1,t_2)\frac{(x_1-t_1)^2}{2}\,dt_2dt_1\end{aligned} \tag{30}$$

and

$$\begin{aligned}&\left|\int_{a_2}^{b_2} f(x_1,t_2)\left(\int_{a_1}^{b_1} w(t_1,t_2)\,dt_1\right)dt_2 - f(x_1,x_2)\int_{a_1}^{b_1}\int_{a_2}^{b_2} w(t_1,t_2)\,dt_2\,dt_1\right. \\ &\qquad \left. + \frac{\partial f}{\partial t_2}(x_1,x_2)\int_{a_1}^{b_1}\int_{a_2}^{b_2}(x_2-t_2)w(t_1,t_2)\,dt_2\,dt_1\right| \le \left\|\frac{\partial^2 f}{\partial t_2^2}\right\|_\infty \int_{a_1}^{b_1}\int_{a_2}^{b_2} w(t_1,t_2)\frac{(x_2-t_2)^2}{2}\,dt_2dt_1.\end{aligned} \tag{31}$$

It is of interest to note that combining (20), (30) and (31) will produce (3). Thus, in one sense, (20) is more general than (3) since it is not obvious how one may derive (20) from (3).

One advantage of (20) over (3) is that the upper bound involves one term instead of three. Thus, with (20) we can find points (x_1,x_2) that will minimize upper bound in terms of the weight and independent of the integrand. In the following corollary we will identify points (x_1,x_2) to minimize the bound

$$\mathcal{J}(x_1,x_2) = \int_{a_1}^{b_1}\int_{a_2}^{b_2} |x_1-t_1|\,|x_2-t_2|\,w\,(t_1,t_2)\,dt_2dt_1. \tag{32}$$

Corollary 10. *$\mathcal{J}(x_1,x_2)$ as defined in (32) is minimized at (x_1^*,x_2^*) where x_1^* and x_2^* satisfy the equations*

$$\int_{a_1}^{x_1^*}\int_{a_2}^{b_2} |x_2^*-t_2|w(t_1,t_2)\;dt_2dt_1 = \int_{x_1^*}^{b_1}\int_{a_2}^{b_2} |x_2^*-t_2|w(t_1,t_2)\;dt_2dt_1 \tag{33}$$

and

$$\int_{a_2}^{x_2^*}\int_{a_1}^{b_1} |x_1^*-t_1|w(t_1,t_2)\;dt_2dt_1 = \int_{x_2^*}^{b_2}\int_{a_1}^{b_1} |x_1^*-t_1|w(t_1,t_2)\;dt_2dt_1. \tag{34}$$

Proof. Evaluating the partial derivatives of $\mathcal{J}$ gives

$$\mathcal{J}^{(1)} = \frac{\partial \mathcal{J}}{\partial x_1}(x_1, x_2) = \int_{a_1}^{x_1}\int_{a_2}^{b_2} |x_2 - t_2| w(t_1, t_2)\, dt_2 dt_1 - \int_{x_1}^{b_1}\int_{a_2}^{b_2} |x_2 - t_2| w(t_1, t_2)\, dt_2 dt_1, \tag{35}$$

$$\mathcal{J}^{(2)} = \frac{\partial \mathcal{J}}{\partial x_2}(x_1, x_2) = \int_{a_2}^{x_2}\int_{a_1}^{b_1} |x_1 - t_1| w(t_1, t_2)\, dt_1 dt_2 - \int_{x_2}^{b_2}\int_{a_1}^{b_1} |x_1 - t_1| w(t_1, t_2)\, dt_1 dt_2. \tag{36}$$

Inspection of (35) reveals that, for fixed x_2, $\mathcal{J}^{(1)}$ is monotonic increasing and $\mathcal{J}^{(1)}(a_1, x_2) = -\mathcal{J}^{(2)}(b_1, x_2) \leq 0$. $\mathcal{J}^{(2)}$ also exhibits similar properties and hence there exists a unique point (x_1^*, x_2^*) that is the zero of (35) and (36) and minimizes $\mathcal{J}$. □

The behaviour of (32) is very dependant on the behaviour of the weight. In Figure 2 contours of $\mathcal{J}$ are plotted for different weight functions. In each case, the minimum point is readily observed and its location depends on the weight and weight null-space.

In the following section, properties of the minimum point of $\mathcal{J}$ are identified for various conditions on w.

4. Minimizing the bound

Solution of equations (33) and (34) provide the point that minimizes the bound (32). The equations are non-linear and two dimensional, thus, in most cases, require numerical treatment. In this section we identify solutions or simplifications to (33) and (34) for specific weight types. Some of these weights are of importance since they appear in the important areas of integral transforms and integral equations.

With functions of two or more variables it is common that an identifiable relationship between the variables is observed. That is, $w(t_1, t_2) = w(\phi(t_1, t_2))$ for some ϕ. For singular weights, the null-space of ϕ, $\{(t_1, t_2) : \phi(t_1, t_2) = 0\}$, may be of interest since this may furnish the singularity structure of the integral. Below, we explore the properties of $\mathcal{J}$ for ϕ being the difference mapping on a square and generalise to more general null-spaces in other corollaries.

Corollary 11 (Difference weight)**.** *Let $w : (a, b) \to (0, \infty)$ be integrable and let $a < x_1, x_2 < b$. Then the bound*

$$\mathcal{J}(x_1, x_2) = \int_a^b \int_a^b |x_1 - t_1|\, |x_2 - t_2|\, w\, |t_1 - t_2|\, dt_2 dt_1.$$

is minimized at the midpoint $x_1 = x_2 = \frac{a+b}{2}$.

Proof. As stated in Corollary 10, $\mathcal{J}$ is minimized at the root of equations (33) and (34). Substituting the midpoint in (33) gives

$$\begin{aligned}
&\int_a^{(a+b)/2}\int_a^b \left|\frac{a+b}{2} - t_2\right| w|t_1 - t_2|\, dt_2 dt_1 - \int_{(a+b)/2}^b\int_a^b \left|\frac{a+b}{2} - t_2\right| w|t_1 - t_2|\, dt_2 dt_1 \\
&\quad = \int_{(a+b)/2}^b\int_a^b \left|\frac{a+b}{2} - v\right| w|u - v|\, dv du - \int_{(a+b)/2}^b\int_a^b \left|\frac{a+b}{2} - t_2\right| w|t_1 - t_2|\, dt_2 dt_1 \\
&\quad = 0,
\end{aligned}$$

where $u = a + b - t_1$ and $v = a + b - t_2$ are integral substitutions. The same treatment on (34) shows that the midpoint minimizes the bound □

The following two corollaries show that the simultaneous equations (33) and (34) may be decoupled under certain conditions for the weight.

Corollary 12 (Separable weight)**.** *Let the conditions in Corollary 10 hold. Furthermore, let w be separable, that is $w(t_1, t_2) = w_1(t_1)w_2(t_2)$, where w_i are themselves weight functions defined on $[a_i, b_i]$, i=1,2. Then $\mathcal{J}$ is minimized at the median of each weight*

$$\int_{a_i}^{x_i} w_i(t_i)\, dt_i = \int_{x_i}^{b_i} w_i(t_i)\, dt_i, \qquad i = 1, 2.$$

Proof. Substituting $w(t_1, t_2) = w_1(t_1)w_2(t_2)$ into (33) and (34) and simplifying produces the result. □

Corollary 13 (Symmetric weight)**.** *Let the conditions in Corollary 10 hold and let $w : (a, b) \times (a, b) \to \mathbb{R}$ be symmetric, that is, $w(t_1, t_2) = w(t_2, t_1)$. Then the minimum point is at $x_1 = x_2$.*

Proof. With the above conditions, the two equations in Corollary 10 are

$$\frac{\partial \mathcal{J}}{\partial x_1}(x_1, x_2) = \int_a^{x_1}\int_a^b |x_2 - t_2| w(t_1, t_2)\, dt_2 dt_1 - \int_{x_1}^b\int_a^b |x_2 - t_2| w(t_1, t_2)\, dt_2 dt_1, \tag{37}$$

$$\frac{\partial \mathcal{J}}{\partial x_2}(x_1, x_2) = \int_a^{x_2}\int_a^b |x_1 - t_1| w(t_1, t_2)\, dt_1 dt_2 - \int_{x_2}^b\int_a^b |x_1 - t_1| w(t_1, t_2)\, dt_1 dt_2. \tag{38}$$

Beginning with (37) we have

$$\begin{aligned}\frac{\partial \mathcal{J}}{\partial x_1}(x_1,x_2) &= \int_a^{x_1}\int_a^b |x_2-t_2|w(t_1,t_2)\,dt_2dt_1 - \int_{x_1}^b\int_a^b |x_2-t_2|w(t_1,t_2)\,dt_2dt_1\\ &= \int_a^{x_1}\int_a^b |x_2-t_1|w(t_2,t_1)\,dt_1dt_2 - \int_{x_1}^b\int_a^b |x_2-t_1|w(t_2,t_1)\,dt_1dt_2\\ &= \int_a^{x_1}\int_a^b |x_2-t_1|w(t_1,t_2)\,dt_1dt_2 - \int_{x_1}^b\int_a^b |x_2-t_1|w(t_1,t_2)\,dt_1dt_2\\ &= \frac{\partial \mathcal{J}}{\partial x_2}(x_2,x_1).\end{aligned}$$

Thus the solution of

$$\frac{\partial \mathcal{J}}{\partial x_1}(x_1,x_2)=0 \qquad \text{and} \qquad \frac{\partial \mathcal{J}}{\partial x_2}(x_1,x_2)=0,$$

is identical to

$$\frac{\partial \mathcal{J}}{\partial x_2}(x_2,x_1)=0 \qquad \text{and} \qquad \frac{\partial \mathcal{J}}{\partial x_2}(x_1,x_2)=0,$$

and hence the solution occurs at $x_1=x_2$. □

In Corollary 11 we showed that if a weight has a "difference" null-space on a square then the bound (32) is minimized at the centre of the square. The following corollary will generalise this result and we will consider a null space of the form $t_1=\phi(t_2)$ where ϕ is anti-symmetric on a rectangle.

Corollary 14. *Let $w:(-a,a)\times(-A,A)\to(0,\infty)$ be a weight function of the form $w(t_1,t_2)=w|t_1-\phi(t_2)|$, where $\phi:(-A,A)\to(-a,a)$ is surjective and odd, for some $a,A>0$, that is $\phi(-t)=-\phi(t)$. Then $\mathcal{J}$ as defined in (32) is minimized at the origin.*

Proof. We need to show that

$$\int_{-a}^0\int_{-A}^A |t_2|w|t_1-\phi(t_2)|\,dt_2dt_1 = \int_0^a\int_{-A}^A |t_2|w|t_1-\phi(t_2)|\,dt_2dt_1 \tag{39}$$

and

$$\int_{-A}^0\int_{-a}^a |t_1|w|t_1-\phi(t_2)|\,dt_1dt_2 = \int_0^A\int_{-a}^a |t_1|w|t_1-\phi(t_2)|\,dt_1dt_2. \tag{40}$$

Making the substitution $t_1=-u$ and $t_2=-v$ in the first integral of (39) we have

$$\int_{-a}^0\int_{-A}^A |t_2|w|t_1-\phi(t_2)|\,dt_2dt_1 = \int_0^a\int_{-A}^A |v|w|u-\phi(v)|\,dvdu.$$

Similarly

$$\int_{-A}^0\int_{-a}^a |t_1|w|t_1-\phi(t_2)|\,dt_1dt_2 = \int_0^A\int_{-a}^a |u|w|u-\phi(v)|\,dudv.$$

Hence, the corollary is proved. □

5. Cubature and grid generation

Theorem 9 can form the basis of a cubature formula for weighted double integrals. That is, we can form a mesh and apply equation (20) to each grid rectangle. The minimum point of each rectangle would be given by (33) and (34). The question that would remain is how would such a grid be "optimally" constructed? For example, for four grid rectangles, as shown in Figure 3, how would ξ_1 and ξ_2 be chosen?

Let us consider a partition $a_i\le\xi_i\le b_i$ of the interval $[a_i,b_i]$, with $x_{i,1}\in[a_i,\xi_i]$ and $x_{i,2}\in[\xi_i,b_i]$, for $i=1,2$. In addition, define D to be the rectangular region $[a_1,b_1]\times[a_2,b_2]$ and define the sub-regions $D_{1,1}=[a_1,\xi_1]\times[a_2,\xi_2]$, $D_{1,2}=[\xi_1,b_1]\times[a_2,\xi_2]$, $D_{2,1}=[a_1,\xi_1]\times[\xi_2,b_2]$ and $D_{2,2}=[\xi_1,b_1]\times[\xi_2,b_2]$. A sketch of this partition is shown is Figure 3.

Theorem 15. *Let the conditions in Theorem 9 hold. Given the partition defined above, the following double integral inequality holds*

$$\begin{aligned}(41)\quad &\left|\iint_D f(t_1,t_2)w(t_1,t_2)\,dt_1dt_2-\sum_{i=1}^{2}\Big(\iint_{D_{1,i}+D_{2,i}} f(x_{1,i},t_2)w(t_1,t_2)\,dt_1dt_2\right.\\ &\left.-\iint_{D_{i,1}+D_{i,2}} f(t_1,x_{2,i})w(t_1,t_2)\,dt_1dt_2\Big)+\sum_{i=1}^{2}\sum_{j=1}^{2} f(x_{1,j},x_{2,i})\iint_{D_{i,j}} w(t_1,t_2)\,dt_1dt_2\right|\\ &\qquad\le\left\|\frac{\partial^2 f}{\partial t_1\partial t_2}\right\|_\infty\sum_{i=1}^{2}\sum_{j=1}^{2}\iint_{D_{i,j}}|x_{1,i}-t_1||x_{2,j}-t_2|w(t_1,t_2)\,dt_1dt_2.\end{aligned}$$

The bound is minimized at the points $x_{i,j},\xi_i$, $(i,j=1,2)$ satisfying

$$\begin{aligned}(42)\quad &\int_{a_1}^{x_{1,1}}\int_{a_2}^{\xi_2}|x_{2,1}-t_2|w(t_1,t_2)\,dt_2dt_1+\int_{a_1}^{x_{1,1}}\int_{\xi_2}^{b_2}|x_{2,2}-t_2|w(t_1,t_2)\,dt_2dt_1\\ &\quad=\int_{x_{1,1}}^{\xi_1}\int_{a_2}^{\xi_2}|x_{2,1}-t_2|w(t_1,t_2)\,dt_2dt_1+\int_{x_{1,1}}^{\xi_1}\int_{\xi_2}^{b_2}|x_{2,2}-t_2|w(t_1,t_2)\,dt_2dt_1,\\ (43)\quad &\int_{\xi_1}^{x_{1,2}}\int_{a_2}^{\xi_2}|x_{2,1}-t_2|w(t_1,t_2)\,dt_2dt_1+\int_{\xi_1}^{x_{1,2}}\int_{\xi_2}^{b_2}|x_{2,2}-t_2|w(t_1,t_2)\,dt_2dt_1\\ &\quad=\int_{x_{1,2}}^{b_1}\int_{a_2}^{\xi_2}|x_{2,1}-t_2|w(t_1,t_2)\,dt_2dt_1+\int_{x_{1,2}}^{b_1}\int_{\xi_2}^{b_2}|x_{2,2}-t_2|w(t_1,t_2)\,dt_2dt_1,\\ (44)\quad &\int_{a_2}^{x_{2,1}}\int_{a_1}^{\xi_1}|x_{1,1}-t_1|w(t_1,t_2)\,dt_1dt_2+\int_{a_2}^{x_{2,1}}\int_{\xi_1}^{b_1}|x_{1,2}-t_1|w(t_1,t_2)\,dt_1dt_2\\ &\quad=\int_{x_{2,1}}^{\xi_2}\int_{a_1}^{\xi_1}|x_{1,1}-t_1|w(t_1,t_2)\,dt_1dt_2+\int_{x_{2,1}}^{\xi_2}\int_{\xi_1}^{b_1}|x_{1,2}-t_1|w(t_1,t_2)\,dt_1dt_2,\\ (45)\quad &\int_{\xi_2}^{x_{2,2}}\int_{a_1}^{\xi_1}|x_{1,1}-t_1|w(t_1,t_2)\,dt_1dt_2+\int_{\xi_2}^{x_{2,2}}\int_{\xi_1}^{b_1}|x_{1,2}-t_1|w(t_1,t_2)\,dt_1dt_2\\ &\quad=\int_{x_{2,2}}^{b_2}\int_{a_1}^{\xi_1}|x_{1,1}-t_1|w(t_1,t_2)\,dt_1dt_2+\int_{x_{2,2}}^{b_2}\int_{\xi_1}^{b_1}|x_{1,2}-t_1|w(t_1,t_2)\,dt_1dt_2,\\ (46)\quad &\xi_1=\frac{x_{1,1}+x_{1,2}}{2}\qquad\textit{and}\qquad\xi_2=\frac{x_{2,1}+x_{2,2}}{2}.\end{aligned}$$

Proof. To obtain (41), it is a simple matter of applying equation (20) of Theorem 9 to each region $D_{i,j}$ $(i,j=1,2)$, summing and finally employing the triangle inequality.

To show equations (42)–(46), we calculate the stationary point of the bound

$$(47)\quad \mathcal{J}=\sum_{i=1}^{2}\sum_{j=1}^{2}\iint_{D_{i,j}}|x_{1,i}-t_1||x_{2,j}-t_2|w(t_1,t_2)\,dt_1dt_2.$$

For $x_{1,1}$,

$$\begin{aligned}\frac{\partial\mathcal{J}}{\partial x_{1,1}}&=\frac{\partial}{\partial x_{1,1}}\left\{\sum_{j=1}^{2}\iint_{D_{1,j}}|x_{1,1}-t_1||x_{2,j}-t_2|w(t_1,t_2)\,dt_1dt_2\right\}\\ &=\frac{\partial}{\partial x_{1,1}}\left\{\int_{a_1}^{x_{1,1}}\int_{a_2}^{\xi_2}(x_{1,1}-t_1)|x_{2,1}-t_2|w(t_1,t_2)\,dt_2dt_1\right.\\ &\qquad+\int_{x_{1,1}}^{\xi_1}\int_{a_2}^{\xi_2}(t_1-x_{1,1})|x_{2,1}-t_2|w(t_1,t_2)\,dt_2dt_1\\ &\qquad+\int_{a_1}^{x_{1,1}}\int_{\xi_2}^{b_2}(x_{1,1}-t_1)|x_{2,2}-t_2|w(t_1,t_2)\,dt_2dt_1\\ &\qquad\left.+\int_{x_{1,1}}^{\xi_1}\int_{\xi_2}^{b_2}(t_1-x_{1,1})|x_{2,2}-t_2|w(t_1,t_2)\,dt_2dt_1\right\}\end{aligned}$$

$$= \int_{a_1}^{x_{1,1}} \int_{a_2}^{\xi_2} |x_{2,1} - t_2| w(t_1, t_2)\, dt_2 dt_1 - \int_{x_{1,1}}^{\xi_1} \int_{a_2}^{\xi_2} |x_{2,1} - t_2| w(t_1, t_2)\, dt_2 dt_1$$
$$+ \int_{a_1}^{x_{1,1}} \int_{\xi_2}^{b_2} |x_{2,2} - t_2| w(t_1, t_2)\, dt_2 dt_1 - \int_{x_{1,1}}^{\xi_1} \int_{\xi_2}^{b_2} |x_{2,2} - t_2| w(t_1, t_2)\, dt_2 dt_1.$$

Setting the last expression to zero gives (42) and the same process can be used to show equations (43)–(45).

To show (46), observe that

$$\frac{\partial \mathcal{J}}{\partial \xi_1} = \int_{a_2}^{\xi_2} (\xi_1 - x_{1,1}) |x_{2,1} - t_2| w(\xi_1, t_2)\, dt_2 - \int_{a_2}^{\xi_2} (x_{1,2} - \xi_1) |x_{2,1} - t_2| w(\xi_1, t_2)\, dt_2$$
$$+ \int_{\xi_2}^{b_2} (\xi_1 - x_{1,1}) |x_{2,2} - t_2| w(\xi_1, t_2)\, dt_2 - \int_{\xi_2}^{b_2} (x_{1,2} - \xi_1) |x_{2,1} - t_2| w(\xi_1, t_2)\, dt_2$$
$$= 2 \int_{a_2}^{\xi_2} \left(\xi_1 - \frac{x_{1,1} + x_{1,2}}{2} \right) |x_{2,1} - t_2| w(\xi_1, t_2)\, dt_2$$
$$+ 2 \int_{\xi_2}^{b_2} \left(\xi_1 - \frac{x_{1,1} + x_{1,2}}{2} \right) |x_{2,2} - t_2| w(\xi_1, t_2)\, dt_2,$$

which obviously has a root at $(46)_1$. Similarly, we can show $(46)_2$. □

We now proceed to a full weighted cubature formulae.

Define the following partitions of the intervals $[a_i, b_i]$

$$I_i : a_i = \xi_{i,0} \le \xi_{i,1} \le \cdots \le \xi_{i,n} = b_i,$$

and let $x_{i,j} \in [\xi_{i,j-1}, \xi_{i,j}]$ for $i = 1, 2$ and $j = 1, 2, \ldots, n$. Furthermore, let $D_{i,j} = [\xi_{1,i-1}, \xi_{1,i}] \times [\xi_{2,j-1}, \xi_{2,j}]$, $D_i^{(1)} = \bigcup_{k=1}^n D_{i,k}$ and $D_i^{(2)} = \bigcup_{k=1}^n D_{k,i}$, for $i, j = 1, 2, \ldots, n$.

Consider the weighted cubature formula

$$A(f, w, I_1, I_2, \boldsymbol{\xi}, \boldsymbol{x}) \tag{48}$$
$$= \sum_{i=1}^n \left(\iint_{D_i^{(1)}} f(x_{1,i}, t_2) w(t_1, t_2)\, dt_1 dt_2 + \iint_{D_i^{(2)}} f(t_1, x_{2,i}) w(t_1, t_2)\, dt_1 dt_2 \right)$$
$$- \sum_{i=1}^n \sum_{j=1}^n f(x_{1,i}, x_{2,j}) \iint_{D_{i,j}} w(t_1, t_2)\, dt_1 dt_2.$$

Using the above assumptions, we can write the following theorem.

Theorem 16. *Let $f : [a_1, b_1] \times [a_2, b_2] \to \mathbb{R}$ and $w : (a_1, b_1) \times (a_2, b_2) \to (0, \infty)$ be as in Theorem 9 and $I_1, I_2, \boldsymbol{\xi}, \boldsymbol{x}$ be given above. The following weighted cubature formula holds*

$$\int_{a_1}^{b_1} \int_{a_2}^{b_2} f(t_1, t_2) w(t_1, t_2)\, dt_2 dt_1 = A(f, w, I_1, I_2, \boldsymbol{\xi}, \boldsymbol{x}) + R(f, w, I_1, I_2, \boldsymbol{\xi}, \boldsymbol{x}), \tag{49}$$

where

$$|R(f, w, I_1, I_2, \boldsymbol{\xi}, \boldsymbol{x})| \le \left\| \frac{\partial^2 f}{\partial t_1 \partial t_2} \right\|_\infty \sum_{i=1}^n \sum_{j=1}^n \iint_{D_{i,j}} |x_{1,i} - t_1| |x_{2,j} - t_2| w(t_1, t_2)\, dt_1 dt_2. \tag{50}$$

The bound (50) is minimized when $\boldsymbol{x}$ and $\boldsymbol{\xi}$ satisfy

$$\sum_{j=1}^n \int_{\xi_{1,i-1}}^{x_{1,i}} \int_{\xi_{2,j-1}}^{\xi_{2,j}} |x_{2,j} - t_2| w(t_1, t_2)\, dt_2 dt_1 = \sum_{j=1}^n \int_{x_{1,i}}^{\xi_{1,i}} \int_{\xi_{2,j-1}}^{\xi_{2,j}} |x_{2,j} - t_2| w(t_1, t_2)\, dt_2 dt_1 \tag{51}$$

$$\sum_{j=1}^n \int_{\xi_{2,i-1}}^{x_{2,i}} \int_{\xi_{1,j-1}}^{\xi_{1,j}} |x_{1,j} - t_1| w(t_1, t_2)\, dt_1 dt_2 = \sum_{j=1}^n \int_{x_{2,i}}^{\xi_{2,i}} \int_{\xi_{1,j-1}}^{\xi_{1,j}} |x_{1,j} - t_1| w(t_1, t_2)\, dt_1 dt_2 \tag{52}$$

$$\xi_{k,l} = \frac{x_{k,l} + x_{k,l+1}}{2}, \qquad \text{for } i = 1, \ldots, n,\ l = 1, \ldots, n-1,\ k = 1, 2. \tag{53}$$

Proof. The proof follows that of Theorem 16. □

To find the $4n - 2$ unknowns

$$x_{i,1} \le \xi_{i,1} \le x_{i,2} \le \cdots \le \xi_{i,n-1} \le x_{i,n},$$

for $i = 1, 2$, we need to solve the $4n - 2$ coupled non-linear equations (51), (52) and (53). These equations are easily solved iteratively with a uniform grid as the starting point. With this method of solution all variables

are fixed apart from the parameter of interest. Thus for example if $k = 1$ and we fix i, then equation (51) may be considered as a function of $x_{1,i}$ only; say $F(x_{1,i})$. It is easy to see that

$$F'(x_{1,i}) = 2\sum_{j=1}^{n} \int_{\xi_{2,j-1}}^{\xi_{2,j}} |x_{2,j} - t_2| w(x_{1,i}, t_2)\, dt_2 \geq 0$$

and $F(\xi_{1,i-1}) \leq 0$, $F(\xi_{1,i}) \geq 0$. Thus F has a unique root and the bisection algorithm would be an appropriate numerical technique to produce the solution.

In Figures 4, 5, 6 and 7, the grid obtained via numerical solution of (51)–(53) is plotted for various weight functions and n. We can see that the grid clustering reflects the weight behaviour.

Acknowledgments

This work was completed while the second named author was on study leave from Victoria University to the School of Mathematics and Statistics, The University of Birmingham, UK. The author acknowledges support provided from both institutions.

References

[1] N.S. Barnett and S.S. Dragomir, An Ostrowski type inequality for double integrals and application for cubature formulae, *Soochow J. Math.*, **27**(1), 1–10 (2001).

[2] G. Hanna, P. Cerone and J. Roumeliotis, An Ostrowski type inequality in two dimensions using the three point rule, *ANZIAM J.*, **42**(E), C671–C689 (2000). [ONLINE] Available from the internet at `http://anziamj.austms.org.au/V42/CTAC99/Hann/home.html`

[3] G.V. Milovanović, On some integral inequalities, *Univ. Beograd. Publ. Elektrotehn. Fak, Ser. Mat. Fiz.*, No. 498–541, 119–124 (1975).

[4] D. S. Mitrinović, J. E. Pečarić, and A. M. Fink, *Inequalities for functions and their integrals and derivatives*, Kluwer Academic, Dordrecht, 1994.

[5] Feng Qi, Inequalities for a weighted multiple integral, *J. Math. Anal. Appl.*, **253**(2), 381–388 (2001).

[6] J. Roumeliotis, P. Cerone and S.S. Dragomir, An Ostrowski type inequality for weighted mappings with bounded second derivative, *J. KSIAM*, **3**(2), 107–119 (1999).

[7] J. Roumeliotis, Product inequalities and weighted quadrature, in *Ostrowski type inequalities and applications in numerical integration*, (S.S. Dragomir and T.M. Rassias, eds.). In press for Kluwer Academic.

[8] A.H Stroud, *Approximation Calculation of Multiple Integrals*, Prentice Hall, New Jersey, 1971.

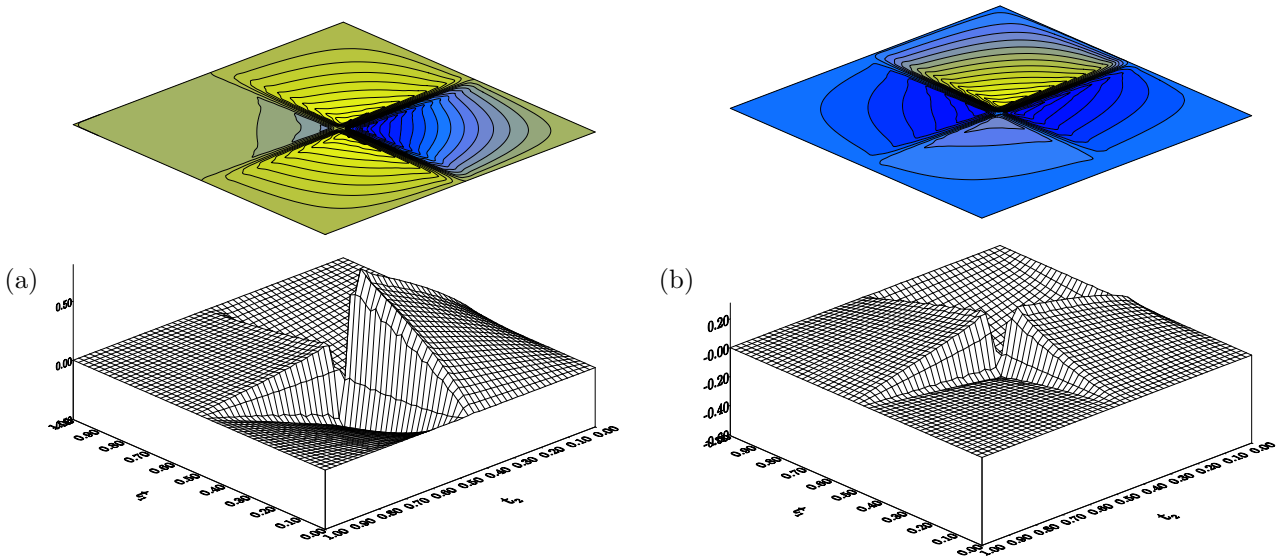


Figure 1. Surface and contour plots of the Peano type kernels P defined in (17) for different weights. (a) $w(t_1, t_2) = -\ln(t_1 t_2)$ over the unit square and $x_1 = x_2 = 0.5$, (b) $w(t_1, t_2) = \sqrt{t_1/t_2}$ over the unit square and $x_1 = x_2 = 0.5$.

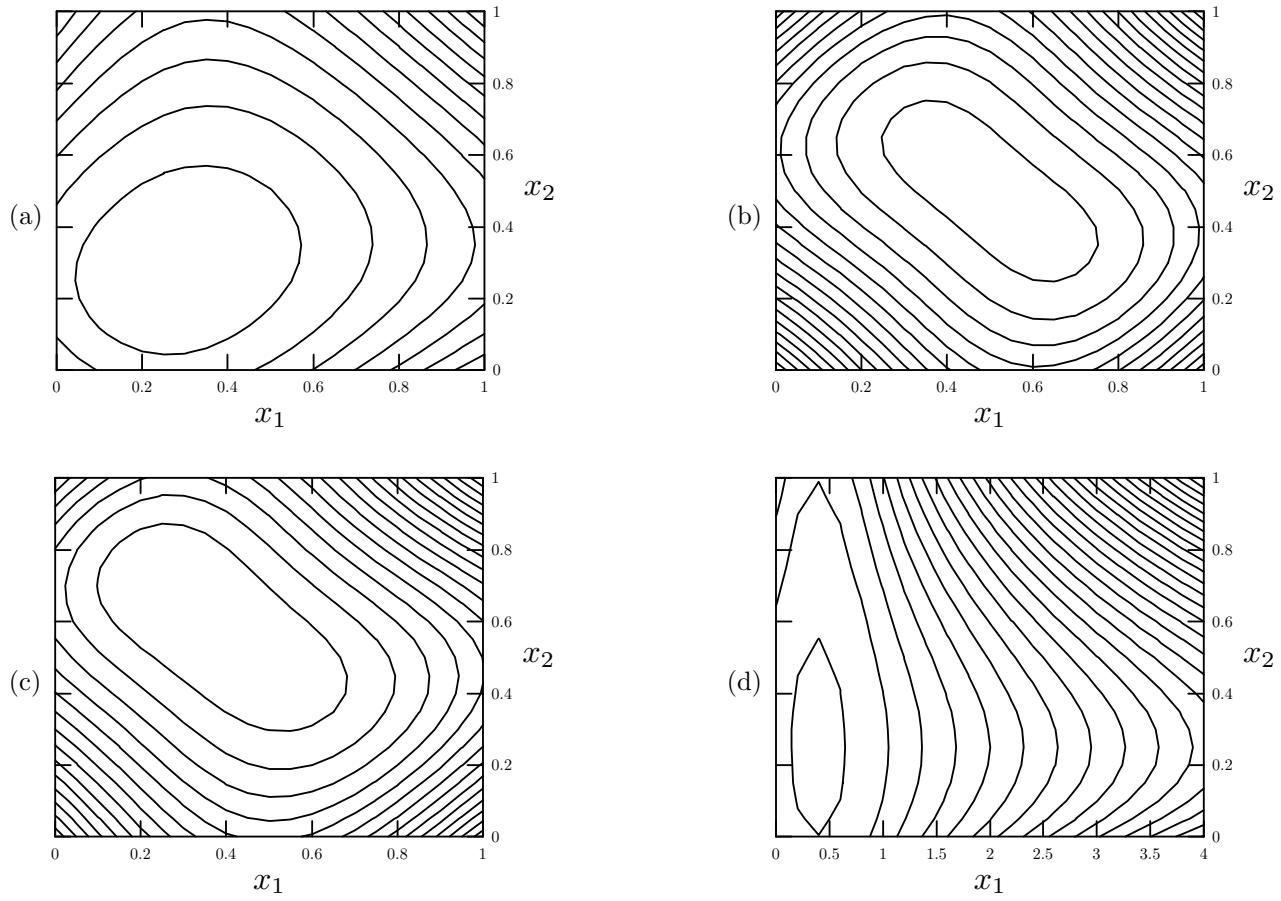


FIGURE 2. Contour plots of the $\mathcal{J}(x_1, x_2)$ given by (32) for various weight functions. (a) $w(t_1, t_2) = -\ln(t_1 t_2)$, $(t_1, t_2) \in (0,1) \times (0,1)$, (b) $w(t_1, t_2) = -\ln|t_1 - t_2|$, $(t_1, t_2) \in (0,1) \times (0,1)$, (c) $w(t_1, t_2) = -\ln|t_1 - t_2^2|$, $(t_1, t_2) \in (0,1) \times (0,1)$ and (d) $w(t_1, t_2) = e^{-t_1}/\sqrt{t_2}$, $(t_1, t_2) \in (0,4) \times (0,1)$.

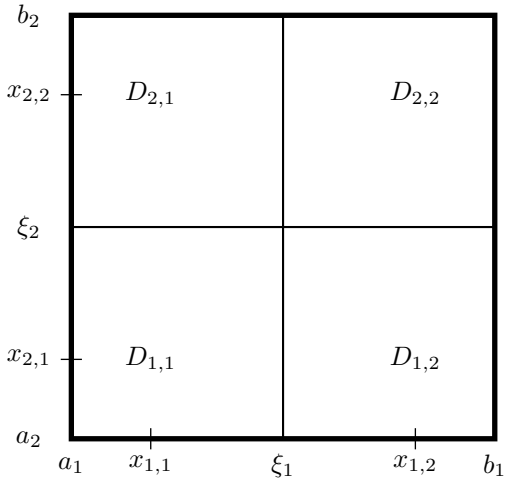


FIGURE 3. A partition of the rectangular region $D = [a_1, b_1] \times [a_2, b_2]$ showing the sub-regions $D_{i,j}$, $i, j = 1, 2$.

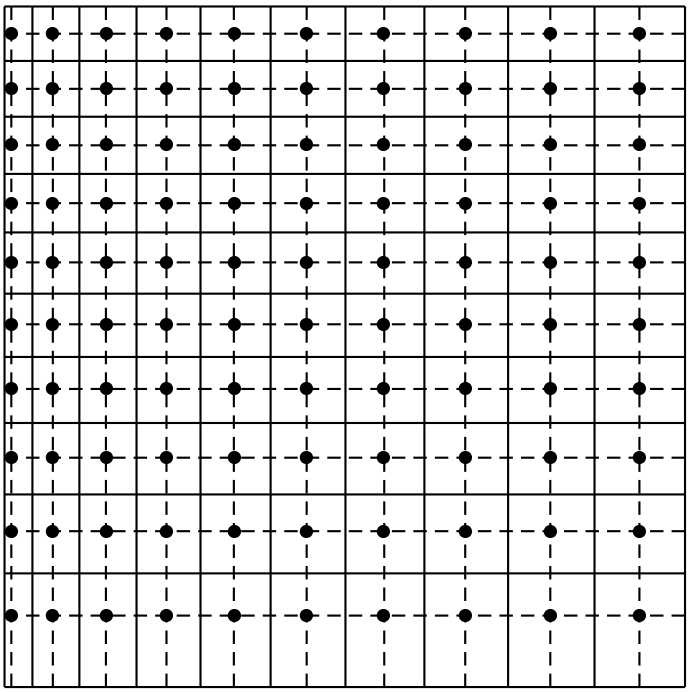


FIGURE 4. Grid generated from the solution of equations (51)– (53) for the weight $w(t_1, t_2) = \sqrt{t_2/t_1}$ over $[0,1] \times [0,1]$ and $n = 10$. The solid lines indicate the composite grid; in each grid square there is one function evaluation (dot) and two single integral evaluations (dashed lines).

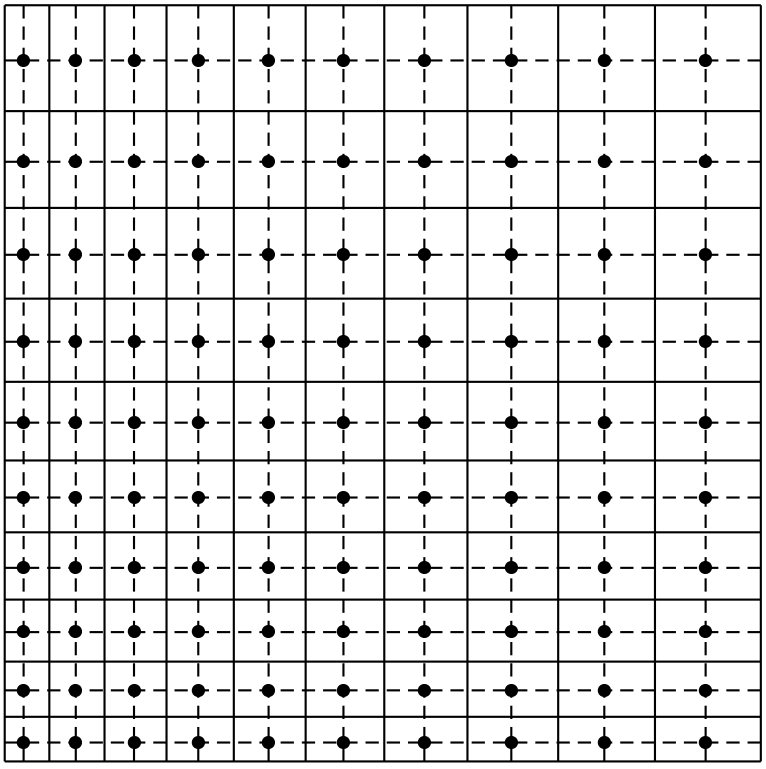


FIGURE 5. Grid generated from the solution of equations (51)– (53) for the weight $w(t_1, t_2) = -\ln(t_1 t_2)$ over $[0,1] \times [0,1]$ and $n = 10$. The solid lines indicate the composite grid; in each grid square there is one function evaluation (dot) and two single integral evaluations (dashed lines).

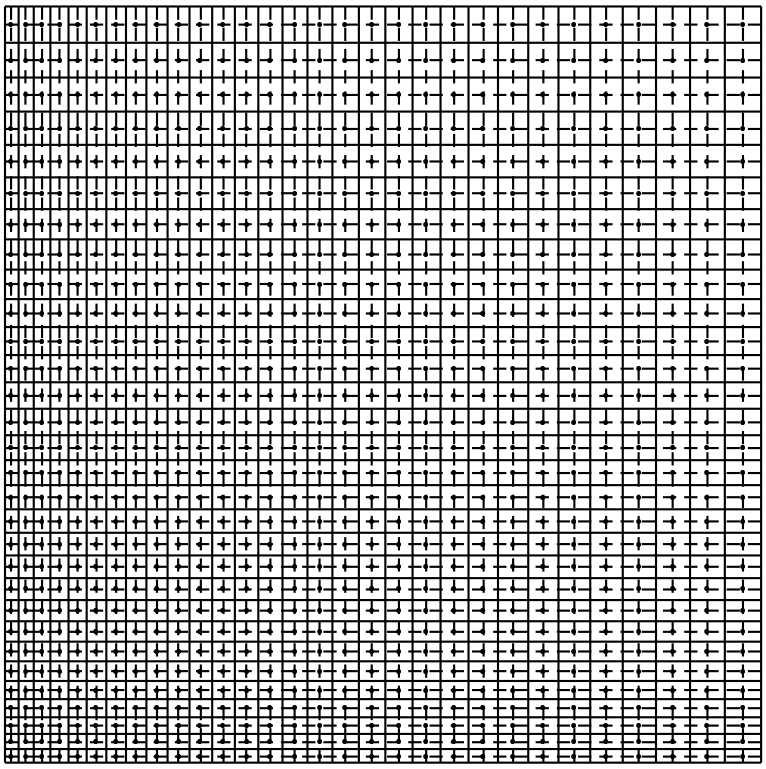


FIGURE 6. Grid generated from the solution of equations (51)– (53) for the weight $w(t_1, t_2) = -\ln(t_1 t_2)$ over $[0, 1] \times [0, 1]$ and $n = 30$. The solid lines indicate the composite grid; in each grid square there is one function evaluation (dot) and two single integral evaluations (dashed lines).

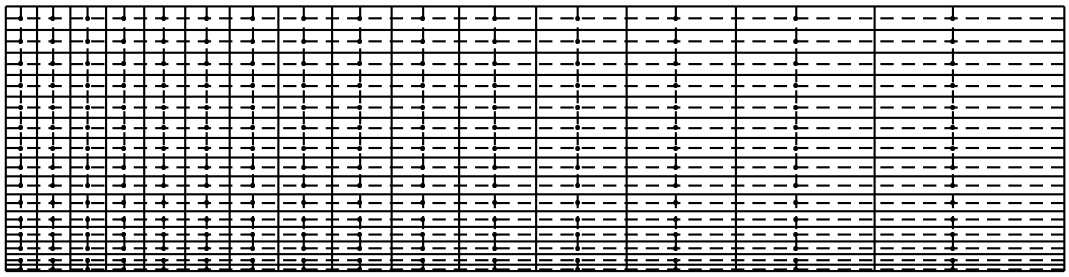


FIGURE 7. Grid generated from the solution of equations (51)– (53) for the weight $w(t_1, t_2) = e^{-t_1}/\sqrt{t_2}$ over $[0, 4] \times [0, 1]$ and $n = 15$. The solid lines indicate the composite grid; in each grid square there is one function evaluation (dot) and two single integral evaluations (dashed lines).

JOURNAL OF CONCRETE AND APPLICABLE MATHEMATICS,VOL.3,NO.4,405-412,
2005,COPYRIGHT 2005 EUDOXUS PRESS,LLC

On a class of polynomials generalizing the Laguerre family

Giuseppe Dattoli

ENEA – Unità di Fisica Teorica, Centro Ricerche Frascati, Via E. Fermi, 45 - 00044 - Frascati (RM), Italia - e-mail: dattoli@frascati.enea.it

Paolo E. Ricci

Università di Roma "La Sapienza", Dipartimento di Matematica, P.le A. Moro, 2 00185 Roma, Italia - e-mail: riccip@uniroma1.it

Clemente Cesarano

ENEA – Unità di Fisica Teorica, Centro Ricerche Frascati, Via E. Fermi, 45 - 00044 - Frascati (RM), Italia & Ulm University, Department of Mathematics - Ulm, Germany - e-mail: cesarano@frascati.enea.it

Abstract

Let $\{p_n(x)\}_{n=0}^{\infty}$ be a sequence of polynomials of degree n, defined by a special class of generating function generalizing the generating function of the Laguerre polynomilas, by using the factorization method, introduced in [1], we determine their differential equations, and some other properties.

2000 Mathematics Subject Classification. 33C45, 34A05.
Key words and phrases. Laguerre polynomials, Appell polynomials, differential equations.

1 Introduction

In recent articles [4], [5], [8], different versions of the factorization method [1] have been used in order to investigate properties of some recurrent polynomials $\{p_n(x)\}_{n=0}^{\infty}$, where $p_n(x)$ is a polynomial of degree n.

G. Dattoli et al. (see [2], [3]) introduced the so called *monomiality principle* as an instrument useful to study in an unified way a large class of special functions, including the multivariable Bessel functions, Hermite and Laguerre polynomials, classical Bernoulli and Euler polynomials and many others.

A family of special polynomials is called *quasi monomial* if there exist two operators $\hat{P}$ and $\hat{M}$, satisfying the commutation property: $[\hat{P}, \hat{M}] := \hat{P}\hat{M} - \hat{M}\hat{P} = 1$, and such that:

$$\hat{P}(p_n(x)) = np_{n-1}(x), \qquad \hat{M}(p_n(x)) = p_{n+1}(x). \tag{1.1}$$

This definition can be extended to some class of special functions, including the multi-variable or multi-indices cases.

From equation (1.1) some important properties of the polynomial family can be easily deduced, such as:

- the differential equation:
$$\hat{M}\hat{P}(p_n(x)) = np_n(x),$$
- the explicit form:
$$p_n(x) = \hat{M}^n(1),$$
- the generating function:
$$e^{t\hat{M}(1)} = \sum_{n=0}^{\infty} \frac{t^n}{n!}\, p_n(x).$$

A more classical approach, connected with the work of L. Infeld and T.E. Hull [1], has been considered by M.X. He and P.E. Ricci [4], [5]. They define two operators L_n^- and L_n^+, depending on n, satisfying the properties:

$$L_n^- p_n(x) = p_{n-1}(x), \qquad L_n^+ p_n(x) = p_{n+1}(x), \tag{1.2}$$

and consequently

$$(L_{n+1}^- L_n^+)p_n(x) = p_n(x), \qquad (L_{n-1}^+ L_n^-)p_n(x) = p_n(x), \tag{1.3}$$

$$(L_{n-k+1}^- L_{n-k+2}^- \cdots L_{n-1}^- L_n^-)p_n(x) = p_{n-k}(x), \tag{1.4}$$

$$(L_{h-1}^+ L_{h-2}^+ \cdots L_1^+ L_0^+)p_0(x) = p_h(x). \tag{1.5}$$

Equations (1.3) give back the differential equation satisfied by the polynomials, and furthermore the formulas (1.4)-(1.5) permit to recover the operator L_n^+ (L_n^-) when the operator L_n^- (L_n^+) together with a linear recurrence relation of any order satisfied by the polynomials $p_n(x)$ is known.

It is worth to note that, the knowledge of the exponential generating function

$$G(x,t) := \sum_{n=0}^{\infty} p_n(x)\frac{t^n}{n!} \tag{1.6}$$

of the given family of polynomials often permits the construction of the operators (1.2). A very useful survey about different classes of generating functions is given in the classical book of H.M. Srivastava and H.L. Manocha [6].

In recent papers, the above methods have been used in order to derive the differential equations satisfied by polynomials belonging to the Appell family [4], [5], [7] and some particular case of d-orthogonal polynomials (see [9], [10], [11]).

Appell polynomials are characterized by the generating function

$$\sum_{n=0}^{\infty} \frac{t^n}{n!} p_n(x) = A\,(t)\, e^{xt}, \qquad p_0(x) = 1, \tag{1.7}$$

with $A(t)$ being a function of t without singularities at $t = 0$. The main characteristic of $A(t)$ is that

$$\frac{d}{dx} p_n(x) = n p_{n-1}(x) \tag{1.8}$$

according to which the ordinary derivative can be viewed as a negative shift operator in the discrete index n.

It has been recently emphasized that polynomials characterized by the generating function [7],[12],[13]

$$G(t, x|0) = \sum_{n=0}^{\infty} \frac{t^n}{n!} l_n(x) = B(t) C_0(xt) \tag{1.9}$$

with $B(t)$ playing the same role of $A(t)$ and

$$C_0(x) = \sum_{r=0}^{n} \frac{(-1)^r x^r}{(r!)^2} \tag{1.10}$$

being the 0^{th} order Tricomi function, are characterized by the property [2]

$$-\frac{d}{dx}\; x\, \frac{d}{dx}\; l_n(x) = n l_{n-1}(x) \tag{1.11}$$

which shares some analogy with the properties of Appell polynomials. Such analogy is reinforced by the fact that the operator

$$\hat{\mathcal{D}}_x := -\frac{d}{dx}\; x\, \frac{d}{dx} \tag{1.12}$$

is recognized as a Laguerre derivative within the context of the monomiality principle applied to Laguerre polynomials [2],[13].

Since for $B(t) = e^t$ the equation (1.9) gives back a family of Laguerre-type polynomials, we will denote by $l_n(x)$ the relevant polynomials.

As a further example we consider the case

$$\begin{aligned} A(t) &= C_m(t) \\ C_m(t) &= \sum_{r=0}^{\infty} \frac{(-1)^r x^r}{r!(n+r)!} \end{aligned} \tag{1.13}$$

where $C_m(t)$ is the m^{th} order Tricomi function, eigenfunction of the operator

$$\hat{\mathcal{D}}_{t,m} = -t\frac{d^2}{dt^2} - (m+1)\frac{d}{dt} \tag{1.14}$$

In this paper we address the problem of obtaining general conditions on the function $B(t)$ to derive the differential equations specifying $l_n(x)$. A first effort to derive differential equations for the Laguerre type family has been put forward in ref. [7], here we will develop a more systematic strategy also employing the methods of refs. [4]-[5].

2 Laguerre type polynomials and relevant differential equations

We have already remarked that the 0^{th} order Tricomi function is an eigenfunction of the Laguerre derivative and that the m^{th} order functions are eigenfunctions of the operator defined by equation (1.5). We will therefore derive the differential equations for families of Laguerre type polynomials by just taking advantage from this fact and consider as introductory example the generating function

$$G(t;x|m) = \sum_{n=0}^{\infty} t^n l_n^{(m)} = \frac{C_m(xt)}{1-t}, \tag{2.1}$$

with

$$l_n^{(m)}(x) = \sum_{r=0}^{n} \frac{(-x)^r}{r!(m+r)!}. \tag{2.2}$$

By applying the operator (1.5) to both sides of

$$(1-t)G(x,t|m) = C_m(xt), \tag{2.3}$$

we find for the generating function

$$t(1-t)\frac{d^2}{dt^2}\, G + [(m-1)t - (m+1)]\, \frac{dG}{dt} + [m+1+x(1-t)]\, G = 0, \tag{2.4}$$

which provides the recursion

$$\begin{aligned} l_{n+1}^{(m)}(x) &= \frac{1}{n-m-1}\left\{ xn l_{n-1}^{(m)}(x) + \right. \\ &+ \left. [n(n-1) - n(m-1) - (m+1) - x]\, l_n^{(m)}(x) \right\}. \end{aligned} \tag{2.5}$$

In this case the second recursion is provided by (see equation (1.5)) and the concluding section)

$$\hat{\mathcal{D}}_{x,m}\; l_n^{(m)}(x) = n l_{n-1}^{(m)}(x) \tag{2.6}$$

so that we can write the differential equation for the $l_n^{(m)}(x)$ polynomials in the form

$$\begin{aligned} x\hat{\mathcal{D}}_{x,m}^2\; l_n^{(m)}(x) &+ [(n-1)(n-2) - (n-1)(m-1) + \\ &+ m - 1 - x]\, \hat{\mathcal{D}}_{x,m}\; l_n^{(m)}(x) = n(n-m-2) l_n^{(m)}(x), \end{aligned} \tag{2.7}$$

to this aim it is worth noting that the Laguerre derivative satisfies the relations [2],[13]

$$\begin{aligned}\hat{\mathcal{D}}_x^n &= (-1)^n \frac{d^n}{dx^n}\, x^n\, \frac{d^n}{dx^n} = \\ &= (-1)^n n! \sum_{k=0}^{n} \binom{n}{k} \frac{x^k}{k!} \left(\frac{d}{dx}\right)^{n+k}.\end{aligned} \tag{2.8}$$

The method we have just outlined can be generalized, to this aim we consider the generating function

$$\Gamma(x,t|m) = \sum_{n=0}^{\infty} \frac{t^n}{n!}\, \lambda_n^{(m)}(x) = B\,(t)\, C_m(xt), \tag{2.9}$$

and on account of the identity

$$\frac{d^s}{dx^s}\, C_n(x) = (-1)^s C_{n+s}(x) \tag{2.10}$$

it is shown that the polynomials $\lambda_n^{(m)}(x)$ satisfy the recurrence

$$\frac{d}{dx}\, \lambda_n^{(m)}(x) = -n\lambda_{n-1}^{(m+1)}(x). \tag{2.11}$$

Limiting ourselves to the case when $m = 0$, we find that the generating function satisfies the equation

$$\begin{aligned}-\frac{d}{dt}\, t\, \frac{d}{dt}\, \Gamma(x,t|0) &= [x + S_0(t) + S_1(t)]\, \Gamma(x,t|0) + \\ &+ 2xtS_0(t)\Gamma\,(x,t|1)\end{aligned} \tag{2.12}$$

where

$$S_0(t) = -\frac{1}{B(t)} \frac{dB(t)}{dt}, \quad S_1(t) = -\frac{1}{B(t)} \frac{d^2B(t)}{dt^2}. \tag{2.13}$$

By assuming that $S_{0,1}(t)$ are known through their series expansion

$$S_{0,1}(t) = \sum_{r=0}^{\infty} \frac{\alpha_r^{(0,1)} t^r}{r!}, \tag{2.14}$$

we obtain for the $\lambda_n^{(0)}(x)$ the following recursion involving the $\lambda_n^{(1)}(x)$

$$\begin{aligned}\lambda_{n+1}^{(0)}(x) &= \frac{1}{n}\left[x\lambda_n^{(0)}(x) + \pi_n^{(0)}(x) + \pi_n^{(1)}(x) + 2xn\gamma_{n-1}(x)\right], \\ \pi_n^{(0,1)}(x) &= \sum_{r=0}^{n} \frac{\alpha_r^{(0,1)}}{r!}\, \hat{\mathcal{D}}_x^r\, \lambda_n^{(0)}(x), \\ \gamma_n(x) &= \sum_{r=0}^{n} \frac{\alpha_r^{(1)}}{r!}\, \hat{\mathcal{D}}_x^r\, \lambda_n^{(1)}(x)\end{aligned} \tag{2.15}$$

which allows the derivation of a rather involuted differential equation satisfied by the $\lambda_n^{(0)}(x)$.

Simpler results can be obtained by using the relation

$$\Gamma(x,t|0)\frac{1}{B(t)} = C_0(xt), \tag{2.16}$$

in this case assuming known the expansion

$$\frac{1}{B(t)} = \sum_{s=0}^{\infty}\beta_s\frac{t^s}{s!} \tag{2.17}$$

we end up with the recursion

$$\begin{aligned} q_{n+1}(x) &= -\frac{x}{n+1}\ q_n, \\ q_n(x) &= \sum_{s=0}^{n}\binom{n}{s}\beta_s\lambda_{n-s}^{(0)}(x), \end{aligned} \tag{2.18}$$

which has the advantage of not involving the polynomials $\lambda_n^{(m)}(x)$ and yields a less involved differential equation.

Further comments will be presented in the concluding section.

3 Concluding remarks

In this paper we have considered Appell and Laguerre type polynomials. Let us remark that, in general, if

$$A(t) = \sum_{s=0}^{n}\frac{a_s t^s}{s!}, \qquad B(t) = \sum_{s=0}^{n}\frac{b_s t^s}{s!} \tag{3.1}$$

we find

$$\begin{aligned} p_n(x) &= \sum_{r=0}^{n}\binom{n}{r}a_{n-r}x^r, \\ l_n(x) &= \sum_{r=0}^{n}\binom{n}{r}\frac{b_{n-r}}{r!}x^r. \end{aligned} \tag{3.2}$$

Furthermore if $A(t)$ and $B(t)$ possess the Fourier transforms $\widetilde{A}(\sigma)$ and $\widetilde{B}(\sigma)$, we find that the polynomials (3.2) are specified by the integral transforms

$$\begin{aligned} p_n(x) &= \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{+\infty}\widetilde{A}(\sigma)(x+i\sigma)^n d\sigma, \\ l_n(x) &= \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{+\infty}\widetilde{B}(\sigma)L_n(x,i\sigma)d\sigma, \end{aligned} \tag{3.3}$$

where $L_n(x,y)$ are the two variable Laguerre polynomials (see ref. [13],[14]), namely

$$L_n(x,y) = n! \sum_{r=0}^{n} \frac{(-x)^r y^{n-r}}{(n-r)!(r!)^2} \tag{3.4}$$

In a forthcoming investigation we will analyze the case of multi-index polynomials and show how the techniques of this paper can be suitable extended to obtain the differential equations specifying multi-index and multi-variable Appell and Laguerre polynomials.

References

[1] L. Infeld and T. E. Hull: The Factorization Method, *Rev. Mod. Phys.*, **23** (1951), 21–68.

[2] G. Dattoli, Hermite-Bessel and Laguerre-Bessel functions: A by-product ot the monomiality principle, in *Advanced Special Functions and Applications* (Proceedings of the Melfi School on Advanced Topics in Mathematics and Physics; Melfi, 9-12 May 1999) (D. Cocolicchio, G. Dattoli and H.M. Srivastava, Editors), Aracne Editrice, Rome, 2000, pp. 147–164.

[3] G. Dattoli, S. Lorenzutta and C. Cesarano: Finite sums and Generalized forms of Bernoulli polynomials, *Rend. Mat. Appl.*, (VII) **19** (1999), 385–391.

[4] M.X. He and P.E. Ricci, Differential Equation of Appell polynomials via the Factorization Method, *J. Comput. Appl. Math.*, **139** (2002), 231–237.

[5] M.X. He and P.E. Ricci, Differential Equations of Some Classes of Special Functions via the Factorization Method, *J. Comput. Anal. Appl.*, **6** (2004).

[6] H.M. Srivastava and H.L. Manocha, *A Treatise on Generating Functions*, Wiley, New York, 1984.

[7] G. Dattoli, C. Cesarano, D. Sacchetti: A Note on Trucated Polynomials, *Appl. Math. Comput.*, **134** (2003), 595–605.

[8] G. Dattoli, P.E. Ricci and C. Cesarano: Differential equations for Appell type polynomials, *Fract. Calc. Appl. Anal.*, **5** (2002), 69–75.

[9] P. Maroni, L'orthogonalité et les récurrences de polynômes d'ordre supérieur à deux, *Ann. Fac. Sci. Toulouse*, **10** (1) (1989), 105–139.

[10] K. Douak, The relation of the d-orthogonal polynomials to the Appell polynomials, *J. Comput. Appl. Math.*, **70** (1996), 279–295.

[11] Y. B. Cheikh and K. Douak, On the Classical d-Orthogonal Polynomials Defined by Certain Generating Functions, I, *Bull. Belg. Math. Soc.*, **7** (2000), 107–124.

[12] L.C. Andrews: *Special Functions for Engineers and applied Mathematicians*, MacMillan, New York, 1985.

[13] G. Dattoli, A. Torre: Operational methods and two-variable Laguerre polynomials, *Atti Acc. Sc. Torino*, **132** (1995), 1–7.

[14] G. Dattoli, H.M. Srivastava, C. Cesarano, The Laguerre and Legendre polynomials from an operational point of view, *Appl. Math. Comput.*, **124** (2001), 117–127.

JOURNAL OF CONCRETE AND APPLICABLE MATHEMATICS,VOL.3,NO.4,413-435,
2005,COPYRIGHT 2005 EUDOXUS PRESS,LLC

Asymptotic Formulae for Positive Linear Operators on Convex Subsets of Banach Spaces.

Francesco Altomare

Department of Mathematics - University of Bari
Via E. Orabona, 4 - 70125 Bari (Italy)
E-mail: altomare@dm.uniba.it

Sabrina Diomede

Department of Economics - University of Bari
Via C. Rosalba, 53 - 70124 Bari (Italy)
E-mail: s.diomede@dse.uniba.it

Abstract

We establish an asymptotic formula for general sequences of positive linear operators acting on spaces of continuous functions defined on convex subsets of (possibly infinite-dimensional) Banach spaces.

Moreover we deduce analogous formulae for the Bernstein-Schnabl operators on bounded convex subsets of separable Banach spaces, as well as for the multidimensional extension of the so-called Szász-Mirakjan operators.

Key words and phrases: Asymptotic formula, positive operator, Bernstein-Schnabl operator.
Mathematics Subject Classification (2000): 41A80, 41A36, 47N99.

1 Introduction

In the recent paper [2], the authors have determined some general conditions under which a sequence $(L_n)_{n\geq 1}$ of positive linear operators acting on a function space E satisfies an asymptotic formula of the type

$$\lim_n \|n(L_n(f) - f) - A(f)\| = 0 \qquad (f \in D(A))$$

with respect to a weighted norm on E, where A is a second order differential operator defined on a suitable subspace $D(A)$ of E of twice continuously differentiable functions.

The space E involved in the previous formula is a function space on some (possibly unbounded) real interval.

In this Note we treat of the more general case in which E is a space of real-valued functions defined on a convex subset K of an arbitrary Banach space.

In such a general setting, the second order differential operator is replaced by an "abstract differential" operator of the form

$$A(f)(x) := \frac{1}{2}T(f''(x))(x) + f'(x)(\beta(x)) + \gamma(x)f(x),$$

where f is a sufficiently smooth function on K, T an operator acting on the space of all bilinear continuous forms on X, β a mapping from K into X, γ a real-valued function on K and $f'(x)$ and $f''(x)$ denote the first and the second Fréchet derivatives of f at $x \in K$.

Our main results established in Section 3, besides having an intrinsic interest of their own, allow us to obtain some asymptotic formulae for several sequences of positive linear operators as, for instance, the Bernstein-Schnabl operators, defined by

$$B_n(f)(x) := \int_K \cdots \int_K f\left(\frac{x_1 + \cdots + x_n}{n}\right) d\mu_x(x_1) \cdots d\mu_x(x_n),$$

where K is a bounded convex subset of a separable Banach space, $(\mu_x)_{x\in K}$ is a family of probability Borel measures on K, and f belongs to a suitable function space on K.

In this special case, our results generalize those of [1] and [8].

We also discuss in detail the finite dimensional case extending, among other things, some results of [9]. As an application, we establish an asymptotic formula for the multidimensional extension of Szász-Mirakjan operators introduced in [10].

2 Notation and preliminary results

Let $(X, \|\cdot\|)$ be a Banach space and K a convex subset of X. For all $x \in X$ we shall consider the following mappings

$$\psi_x : y \in K \mapsto y - x \in X; \tag{2.1}$$

$$\psi_x^{(2)} : y \in K \mapsto (y - x, y - x) \in X^2; \tag{2.2}$$

$$d_x : y \in K \mapsto \|y - x\| \in \mathbf{R}. \tag{2.3}$$

Moreover, the symbols $\mathcal{F}(X,\mathbf{R})$, $\mathcal{F}(K,X)$, X' and $\mathcal{L}^{(2)}(X,\mathbf{R})$ will stand respectively for the space of all real functions defined on X, the space of all mappings from K into X, the space of all linear and continuous forms on X and, finally, the space of all bilinear and continuous forms on X.

The spaces X' and $\mathcal{L}^{(2)}(X,\mathbf{R})$ will be equipped with the natural norms

$$\|u\| := \|u\|_{X'} := \sup\{|u(x)| \ : x \in X, \|x\| \leq 1\} \quad (u \in X') \tag{2.4}$$

and

$$\|v\| := \|v\|_{\mathcal{L}^{(2)}(X,\mathbf{R})} := \sup\{|v(x,y)| \ : x,y \in X, \|x\| \leq 1, \|y\| \leq 1\} \tag{2.5}$$

$$(v \in \mathcal{L}^{(2)}(X,\mathbf{R})).$$

In the sequel we shall also consider the linear subspace of $\mathcal{F}(K,X)$

$$S(K,X) := span(\{\psi_x \ : \ x \in K\}) \tag{2.6}$$

spanned by the set $\{\psi_x \ : \ x \in K\}$.

Let us now introduce the function spaces in which the next theorems are mainly set.

By $\mathcal{C}(K,\mathbf{R})$ we denote the space of all continuous real-valued functions defined on K; moreover, $\widetilde{\mathcal{C}}^2(K,\mathbf{R})$ stands indifferently for the space

$$\begin{aligned}\mathcal{C}^{(2)}(K,\mathbf{R}) : \ &= \ \{f \in \mathcal{C}(K,\mathbf{R}) : f \text{ is the restriction to K of a twice} \\ &\qquad \text{continuously differentiable function defined on an open} \\ &\qquad \text{neighborhood of K}\}\end{aligned}$$

if the interior $Int(K)$ of K is empty, or for the space

$$\begin{aligned}\mathcal{C}^2_e(K,\mathbf{R}) : \ &= \ \{f \in \mathcal{C}(K,\mathbf{R}) : f \text{ is twice continuously differentiable on } Int(K) \\ &\qquad \text{and its derivatives admit continuous extensions up to } K\}\end{aligned}$$

if $Int(K) \neq \emptyset$. Thus, in this second case, for any $f \in \widetilde{\mathcal{C}}^2(K,\mathbf{R})$ by the symbols f' and f'' we will mean the continuous extensions to K of the derivatives of f in $Int(K)$.

If X is finite dimensional, K is bounded and the boundary ∂K of K is sufficiently smooth, then $\mathcal{C}^{(2)}(K,\mathbf{R}) \subset \mathcal{C}^2_e(K,\mathbf{R})$.

We also set:

$$\widetilde{\mathcal{C}^2}_b(K,R) := \{f \in \widetilde{\mathcal{C}^2}(K,R) \ : \ \sup_{x \in K} \|f''(x)\|_{\mathcal{L}^{(2)}(X,\mathbf{R})} < +\infty\}, \tag{2.7}$$

and

$$U\widetilde{\mathcal{C}^2}_b(K,R) := \{f \in \widetilde{\mathcal{C}^2}_b(K,R) \ : \ f'' \text{ is uniformly continuous on } K\}. \tag{2.8}$$

Let E be a linear subspace of $\mathcal{F}(K,\mathbf{R})$ such that

$$\{f \in \mathcal{C}(K,\mathbf{R}) : \sup_{x\in K} \frac{|f(x)|}{1+\|x\|^2} < +\infty\} \subset E. \tag{2.9}$$

By virtue of (2.9), for any $u \in X'$, $v \in \mathcal{L}^{(2)}(X,\mathbf{R})$ and $x \in X$,

$$u \circ \psi_x \in E \qquad and \qquad v \circ \psi_x^{(2)} \in E. \tag{2.10}$$

For every $f \in \widetilde{\mathcal{C}^2}(K,R)$, $x, y \in K$ set

$$\omega_f(x,y) := \begin{cases} 0 & if\, x = y, \\ \frac{f(y)-f(x)-f'(x)(y-x)-\frac{1}{2}f"(x)(y-x,y-x)}{\|y-x\|^2} & if\, x \neq y \end{cases} \tag{2.11}$$

In what follows, unless otherwise stated, we shall consider a function $f \in \widetilde{\mathcal{C}^2}_b(K,R)$ and set

$$M := sup_{x\in K}\|f''(x)\|_{\mathcal{L}^{(2)}(X,\mathbf{R})}.$$

First notice that necessarily $f \in E$ and $\omega_f(x,\cdot)d_x^2 \in E$; indeed, if $Int(K) \neq \emptyset$, pick $x_0 \in Int(K)$; for any $x \in Int(K)$, by the Taylor's formula with integral remainder (see for instance [6], § 4.3.1)

$$f(x) = f(x_0) + f'(x_0)(x-x_0) + \int_0^1 (1-t)f''(x+t(x-x_0))(x-x_0,x-x_0)dt,$$

and thus

$$|f(x)| \leq |f(x_0)| + \|f'(x_0)\| \parallel x - x_0 \parallel + \frac{M}{2} \parallel x - x_0 \parallel^2; \tag{2.12}$$

by continuity this last inequality may be extended to K and therefore, by (2.9), $f \in E$. This, in turn, together with (2.10), implies that for any $x \in K$ $\omega_f(x,\cdot)d_x^2 \in E$. By analogous reasonings one may deduce that $f \in E$ and $\omega_f(x,\cdot)d_x^2 \in E$ for any $x \in K$ also in the case $Int(K) = \emptyset$.

Moreover, it will be useful to notice that the function $x \in K \mapsto \|f'(x)\| \in \mathbf{R}$ is itself an element of E, too; in fact, if $Int(K) \neq \emptyset$ we may choose $x_0 \in Int(K)$ and then, as a consequence of a mean-value theorem ([6], Corollary 2.1), for any $x \in K$

$$\|f'(x)\| \leq \|f'(x_0) \parallel + M \parallel x - x_0 \parallel; \tag{2.13}$$

by continuity the same inequality holds for any $x \in K$. One may argue analogously to reach the same conclusion if $Int(K) = \emptyset$.

Assume now $Int(K) \neq \emptyset$; for every $x, y \in Int(K)$ by the definition (2.11) and by using the Taylor's formula with integral remainder we deduce that

$$\begin{aligned} \omega_f(x,y)d_x^2(y) &= \int_0^1 (1-t)f''(x+t(y-x))(y-x,y-x)dt + \\ &- \frac{1}{2}f''(x)(y-x,y-x) = \\ &= \int_0^1 (1-t)(f''(x+t(y-x)) - f''(x))(y-x,y-x)dt. \end{aligned} \tag{2.14}$$

Notice that the function $\omega_f(x,y)d_x^2(y)$ is continuous with respect to both x and y, and therefore, by the continuity of f'', and by the Lebesgue's dominated convergence theorem, the equality (2.14) holds also true for any $x, y \in K$.

The previous formula implies that

$$\mid \omega_f(x,y) \mid d_x^2(y) \leq 2M \parallel y-x \parallel^2 \int_0^1 (1-t)dt = M \parallel y-x \parallel^2,$$

and therefore for any $x, y \in K$

$$|\omega_f(x,y)| \leq M. \tag{2.15}$$

Finally notice that if, in addition, f'' is uniformly continuous, then

$$\lim_{y \to x} \omega_f(x,y) = 0 \qquad \text{uniformly w.r.t. } x \in K. \tag{2.16}$$

Indeed, for a given $\varepsilon > 0$, there exists $\delta > 0$ such that $\|f''(x) - f''(y)\| \leq \varepsilon$ for any $x, y \in K$ for which $\|x-y\| < \delta$; therefore, for any such x and y, and by virtue of (2.14) one has

$$\mid \omega_f(x,y) \mid d_x^2(y) \leq \varepsilon \parallel y-x \parallel^2 \int_0^1 (1-t)dt = \frac{\varepsilon}{2} \parallel y-x \parallel^2,$$

whence it follows $\mid \omega_f(x,y) \mid \leq \frac{\varepsilon}{2}$.

3 Asymptotic formulae on convex subsets of Banach spaces.

In this section we establish an asymptotic formula for an arbitrary sequence of positive linear operators. The particular case in which these operators are associated with a family of measures is considered as well. Finally we develop similar results in finite dimensional settings.

As in the previous Section we shall consider a Banach space X, a convex subset K of X and a linear subspace E of $\mathcal{F}(X, \mathbf{R})$ satisfying (2.9).

Theorem 3.1 *Let $(L_n)_{n \in \mathbf{N}}$ be a sequence of positive linear operators from E into $\mathcal{F}(K, \mathbf{R})$, and consider a divergent sequence $(\varphi(n))_{n \in \mathbf{N}}$ of positive integers. Given a mapping*

$$T : \mathcal{L}^{(2)}(X, \mathbf{R}) \to \mathcal{F}(K, \mathbf{R}),$$

w, $\gamma \in \mathcal{F}(K, \mathbf{R})$ and $\beta \in \mathcal{F}(K, X)$, consider the operator $A : \widetilde{\mathcal{C}}^2(K, R) \mapsto \mathcal{F}(K, \mathbf{R})$ defined by setting for any $f \in \widetilde{\mathcal{C}}^2(K, R)$ and $x \in K$

$$A(f)(x) := \frac{1}{2} T(f''(x))(x) + f'(x)(\beta(x)) + \gamma(x) f(x) \tag{3.1}$$

and assume that

(i) for k=0,2

$$\lim_n \|x\|^k w(x)[\varphi(n)(L_n(\mathbf{1})(x)-1)-\gamma(x)] = 0 \qquad \textit{uniformly w.r.t. } x \in K;$$

(ii) for every $n \in \mathbf{N}$ there exists a linear operator $\widetilde{L}_n : S(K,X) \mapsto \mathcal{F}(K,X)$ such that for every $x \in K$ and $u \in X'$

$$L_n(u \circ \psi_x) = u \circ \widetilde{L}_n(\psi_x) \tag{3.2}$$

and for k=0,1

$$\lim_n \|x\|^k w(x)[\varphi(n)\widetilde{L}_n(\psi_x)(x) - \beta(x)] = 0 \qquad \textit{uniformly w.r.t. } x \in K;$$

(iii) for every bounded subset $B \subset \mathcal{L}^{(2)}(X,\mathbf{R})$

$$\lim_n w(x)[\varphi(n)(L_n(v \circ \psi_x^{(2)}))(x) - T(v)(x)] = 0$$

uniformly w.r.t. $x \in K$ and $v \in B$.

Then

(1) for any $f \in \widetilde{\mathcal{C}}_b^2(K,R)$ such that

$$\lim_n w(x)[\varphi(n)(L_n(\omega_f(x,\cdot)d_x^2))(x)] = 0 \textit{ uniformly w.r.t. } x \in K \tag{3.3}$$

one gets

$$\lim_n w[\varphi(n)(L_n(f) - f) - A(f)] = 0 \tag{3.4}$$

uniformly on K.

(2) If

$$k_0 := \sup_{x \in K, n \geq 1} w(x)\varphi(n)L_n(d_x^2)(x) < +\infty, \tag{3.5}$$

and if there exists $q \in]2,+\infty[$ such that $d_x^q \in E$ for all $x \in K$ and

$$\lim_n w(x)\varphi(n)L_n(d_x^q)(x) = 0 \textit{ uniformly w.r.t. } x \in K, \tag{3.6}$$

then for every $f \in U\widetilde{\mathcal{C}}_b^2(K,R)$ formula (3.3), and therefore also formula (3.4), hold true.

Proof. (1). Consider $f \in \widetilde{\mathcal{C}}_b^2(K,R)$; by (2.11) we deduce that

$$f = f(x)\mathbf{1} + f'(x) \circ \psi_x + \frac{1}{2}f''(x) \circ \psi_x^2 + \omega_f(x,\cdot)d_x^2, \tag{1}$$

whence we obtain that for every $n \geq 1$

$$L_n(f) = f(x)L_n(\mathbf{1}) + L_n(f'(x) \circ \psi_x) + \frac{1}{2}L_n(f''(x) \circ \psi_x^2) + L_n(\omega_f(x,\cdot)d_x^2).$$

Then, taking (3.2) into account,

$$\begin{aligned}
&| w(x)[\varphi(n)(L_n(f)-f)(x)-A(f)(x)] |\leq \\
&\qquad \leq \ |w(x)f(x)[\varphi(n)(L_n(\mathbf{1})(x)-1)-\gamma(x)] \,| + \\
&\qquad + \ | w(x)(\varphi(n)[L_n(f'(x)\circ\psi_x)(x)-(f'(x)\circ\beta)(x)] \,| + \\
&\qquad + \ | w(x)[\frac{\varphi(n)}{2}L_n(f''(x)\circ\psi_x^2)(x)-\frac{1}{2}T(f"(x))(x)] \,| + \\
&\qquad + \ | w(x)\varphi(n)L_n(\omega_f(x,\cdot)d_x^2)(x) \,|\leq \\
&\qquad \leq \ | w(x)f(x)[\varphi(n)(L_n(\mathbf{1})(x)-1)-\gamma(x)] \,| + \\
&\qquad + \ \| f'(x) \,\| \ \|w(x)(\varphi(n)\widetilde{L}_n(\psi_x)(x)-\beta(x))\| + \\
&\qquad + \ \frac{1}{2}|w(x)(\varphi(n)L_n(f''(x)\circ\psi_x^2)(x)-T(f''(x))(x))| + \\
&\qquad + \ | w(x)\varphi(n)L_n(\omega_f(x,\cdot)d_x^2)(x) \,| \,.
\end{aligned}$$

By virtue of hypothesis (i), inequality (2.12), and the fact that $\|x\| \leq 1+\|x\|^2$ for all $x \in K$ one obtains

$$\lim_n w(x)f(x)[\varphi(n)(L_n(\mathbf{1})(x)-1)-\gamma(x)] = 0 \tag{2}$$

uniformly on K.
Moreover, combining (2.13) and condition (ii) we deduce that

$$\lim_n \|f'(x)\| \, \|w(x)(\varphi(n)\widetilde{L}_n(\psi_x)(x)-\beta(x))\| = 0 \tag{3}$$

uniformly on K.

On the other hand, as f'' is assumed to be bounded on K, the set $B(f) := \{f''(x) \in \mathcal{L}^{(2)}(X,\mathbf{R}) \,:\, x \in K\}$ is a bounded subset of the space $\mathcal{L}^{(2)}(X,\mathbf{R})$, and thus condition (iii) applies, implying that

$$\lim_n w(x)[\varphi(n)(L_n(f''(x)\circ\psi_x^2)-T(f''(x)))(x)] = 0 \tag{4}$$

uniformly on K. Therefore, taking (3.3), (2), (3) and (4) into account we obtain the assertion.

(2). Consider $f \in U\widetilde{\mathcal{C}}_b^2(K,R)$, and set again $M := sup_{x\in K}\|f''(x)\|$; we shall prove that (3.3) is fulfilled.
To this aim, set first $\varepsilon > 0$; as a consequence of (2.16) there exists $\delta > 0$ such that for every $x, y \in K$, $\|x-y\| \leq \delta$, one gets

$$|\omega_f(x,y)| \leq \frac{\varepsilon}{2k_0}; \tag{5}$$

furthermore, (3.6) implies the existence of some $\nu \in \mathbf{N}$ for which

$$| w(x)\varphi(n)L_n(d_x^q)(x) | \leq \frac{\varepsilon}{2(M+1)}\delta^{q-2} \tag{6}$$

for every $x \in K$ and $n \in \mathbf{N}$, $n \geq \nu$.

Now set $x, y \in K$; if $\|x-y\| \leq \delta$, then

$$|\omega_f(x,y)|d_x^2(y) \leq \frac{\varepsilon}{2k_0}d_x^2(y)$$

whereas, if $\|x-y\| > \delta$,

$$|\omega_f(x,y)|d_x^2(y) \leq Md_x^2(y) \leq M\frac{d_x^q(y)}{\delta^{q-2}}.$$

In any case, for all $x \in K$

$$|\omega_f(x,\cdot)|d_x^2 \leq \frac{\varepsilon}{2k_0}d_x^2 + M\frac{d_x^q}{\delta^{q-2}}. \tag{7}$$

Consider $n \in \mathbf{N}$, $n \geq \nu$, and $x \in K$; by (3.5), (6) and (7) we deduce that

$$\begin{aligned} |w(x)\varphi(n)L_n(\omega_f(x,\cdot)d_x^2)(x)| \leq \\ \leq \ & | w(x) | \varphi(n)L_n(|\omega_f(x,\cdot)|d_x^2)(x) \leq \\ \leq \ & | w(x) | \varphi(n)[\frac{\varepsilon}{2k_0}L_n(d_x^2)(x) + \frac{M}{\delta^{q-2}}L_n(d_x^q)(x)] \leq \varepsilon. \end{aligned}$$

□

In the following Corollary we shall apply Theorem 3.1 to a sequence of linear positive operators defined by means of a family $(\mu_{x,n})_{x\in K,n\geq 1}$ of Borel measures on K such that the linear subspace E introduced in (2.9) also satisfies the inclusion

$$E \subset \bigcap_{x\in K,n\geq 1} \mathcal{L}^1(K,\mu_{x,n}), \tag{3.7}$$

where each $\mathcal{L}^1(K,\mu_{x,n})$ denotes the space of all $\mu_{x,n}$-integrable functions on K.

In the sequel, if $x \in X$ and $\delta > 0$, we set

$$B(x,\delta) := \{y \in X \ : \| \ x-y \ \| < \delta\}.$$

Corollary 3.2 *Let E be a linear subspace of $\mathcal{F}(X,\mathbf{R})$ satisfying (2.9) and (3.7), where $(\mu_{x,n})_{x\in K,n\geq 1}$ is a given family of Borel measures on K, and let $w \in \mathcal{F}(X,\mathbf{R})$. Moreover, consider a divergent sequence of positive integers $(\varphi(n))_{n\geq 1}$ and the sequence $(L_n)_{n\geq 1}$ of positive linear operators from E into $\mathcal{F}(K,\mathbf{R})$ defined by*

$$L_n(f)(x) := \int_K f \, d\mu_{x,n} \qquad (f \in E, x \in K, n \geq 1), \tag{3.8}$$

and assume that it satisfies condition (3.5) of Theorem 3.1, namely

$$k_0 := \sup_{x\in K, n\geq 1} w(x)\varphi(n)L_n(d_x^2)(x) < +\infty;$$

moreover, suppose that for any $\delta > 0$

$$\lim_n \varphi(n)w(x) \int_{K\setminus B(x,\delta)} d_x^2 \, d\mu_{x,n} = 0 \tag{3.9}$$

uniformly w.r.t. $x \in K$.

Then, for every $f \in U\widetilde{\mathcal{C}}_b^2(K,R)$

$$\lim_n w(x)[\varphi(n)(L_n(\omega_f(x,\cdot)d_x^2))(x)] = 0 \text{ uniformly w.r.t. } x \in K$$

and therefore, if also (i), (ii) and (iii) of Theorem 3.1 are fulfilled, the asymptotic formula (3.4) holds true.

Proof. It suffices to adapt the proof of Theorem 2 of [2] to the present case.

□

Remark 3.3 Note that, if each $\mu_{x,n}$ is finite ($x \in K, n \geq 1$), then formula (3.2) holds true. Indeed, as $\|\psi_x\| \leq 1 + d_x^2$ for any $x \in K$, $\int_K \|\psi_x\| d\mu_{t,n} < +\infty$, and thus ψ_x is Bochner-integrable with respect to each $\mu_{t,n}$. In order to obtain (3.2) it suffices to consider the linear operator $\widetilde{L}_n : S(K,X) \to \mathcal{F}(K,X)$ defined by

$$\widetilde{L}_n(\varphi)(t) := \int_K \varphi d\mu_{t,n} \qquad (n \geq 1, \varphi \in S(K,X), t \in K).$$

□

In what follows we apply Corollary 3.2 to obtain an asymptotic formula for a particular sequence of operators of the form (3.8).

Consider a Banach space X and a probability Borel measure μ on X such that

$$\int_X \|x\|^2 \, d\mu(x) < +\infty \tag{3.10}$$

and for every $u \in X'$

$$\int_X u(x) \, d\mu(x) = 0. \tag{3.11}$$

Observe that $\int_X \|x\| \, d\mu(x) < +\infty$, as for any $x \in X$, $\|x\| \leq 1 + \|x\|^2$; moreover for all $v \in \mathcal{L}^{(2)}(X, \mathbf{R})$ the mapping $x \mapsto v(x,x)$ is μ-integrable, because it is continuous and for all $x \in X \mid v(x,x) \mid \leq \| v \| \, \|x\|^2$.

Consider the space

$$E := \{f \in \mathcal{C}(X, \mathbf{R}) \,:\, \sup_{x \in X} \frac{|f(x)|}{1 + \|x\|^2} < +\infty\},$$

which is included in $\mathcal{L}^1(X, \mu)$ by assumption (3.10).

Moreover, for every $f \in E$, $x \in X$ and $n \geq 1$ the function $y \mapsto f(x + \frac{y}{\sqrt{n}})$ is μ-integrable too; indeed, set $M := \sup_{x \in X} \frac{|f(x)|}{1+\|x\|^2}$; then

$$|f(x + \frac{y}{\sqrt{n}})| \leq M(1 + \| x + \frac{y}{\sqrt{n}} \|^2) \leq M(1 + \| x \|^2 + \frac{2}{\sqrt{n}}\|x\|\|y\| + \frac{1}{n}\|y\|^2).$$

Let us now introduce the sequence of positive linear operators from E into $\mathcal{F}(X, \mathbf{R})$ defined by

$$L_n(f)(x) := \int_X f(x + \frac{y}{\sqrt{n}})\, d\mu(y) \qquad (f \in E, x \in X, n \geq 1). \quad (3.12)$$

Corollary 3.4 *For every $f \in U\widetilde{C}_b^2(X)$*

$$\lim_n n(L_n(f)(x) - f(x)) = \frac{1}{2} \int_X f''(x)(y, y) d\mu(y) \quad (3.13)$$

uniformly with respect to $x \in X$.

Proof. We shall apply Corollary 3.2 to the special case in which $K = X$, $w := \mathbf{1}$ and $\varphi(n) := n$ for any $n \geq 1$; to this purpose we remark that the n-th operator in (3.12) may be regarded as a particular case of the one introduced in (3.8) if $\mu_{x,n}$ denotes the image measure of μ under the continuous mapping $\sigma_{x,n}(y) := x + \frac{y}{\sqrt{n}}$ $(y \in X)$.

First observe that for all $n \geq 1$ $L_n(\mathbf{1}) = \mathbf{1}$, and therefore condition (i) of Theorem 3.1 is fulfilled by setting $\gamma = \mathbf{0}$. Consider now $u \in X'$ and $x \in X$; then, taking (3.11) into account, for every $z \in X$ we have

$$L_n(u \circ \psi_x)(z) = \int_X u(z - x + \frac{y}{\sqrt{n}}) d\mu(y) = u(z - x) = u(\psi_x(z)),$$

and thus (3.2) is satisfied by denoting with $\widetilde{L}_n$'s the identity operator on the space $S(X, X)$. Therefore, condition (ii) of Theorem 3.1 holds true for such $\widetilde{L}_n$ and $\beta := \mathbf{0}$.

Next notice that, as for every $v \in \mathcal{L}^{(2)}(X, \mathbf{R})$, $n \geq 1$ and $x \in X$

$$L_n(v \circ \psi_x^{(2)})(x) = \int_X v(\frac{y}{\sqrt{n}}, \frac{y}{\sqrt{n}}) d\mu(y) = \frac{1}{n} \int_X v(y, y) d\mu(y),$$

also condition (iii) of the previously quoted Theorem 3.1 applies if we introduce the operator $T : \mathcal{L}^{(2)}(X, \mathbf{R}) \to \mathcal{F}(X, \mathbf{R})$ defined as

$$T(v)(x) := \int_X v(y, y) d\mu(y) \qquad (x \in X).$$

We now show that the hypothesis of the preceding Corollary are satisfied; firstly we notice that

$$L_n(d_x^2)(x) = \frac{1}{n} \int_X \|y\|^2 \, d\mu(y)$$

for any $x \in X$; secondly for every $\delta > 0$, $n \geq 1$ and $x \in X$ we have

$$\begin{aligned} n \int_{X \setminus B(x,\delta)} d_x^2(y) \, d\mu_{x,n}(y) = \\ &= n \int_X \mathbf{1}_{X \setminus B(x,\delta)}(x + \frac{y}{\sqrt{n}}) \frac{\|y\|^2}{n} d\mu(y) = \\ &= \int_X \mathbf{1}_{X \setminus B(0,\sqrt{n}\delta)}(y) \|y\|^2 d\mu(y) = \int_{X \setminus B(0,\sqrt{n}\delta)} \|y\|^2 d\mu(y), \end{aligned}$$

where $\mathbf{1}_{X \setminus B(x,\delta)}$ and $\mathbf{1}_{X \setminus B(0,\sqrt{n}\delta)}$ denote respectively the characteristic function of $X \setminus B(x, \delta)$ and $X \setminus B(0, \sqrt{n}\delta)$.

From the above equalities, and by virtue of hypothesis (3.10), we may apply Corollary 3.2 to the present case and then (3.13) holds true.

□

Let us turn our attention to the case in which K is a convex subset of $\mathbf{R}^p$, where $p \geq 1$. The space $\mathbf{R}^p$ will be endowed with an arbitrary norm $\| \cdot \|$ and, in analogy with the notation introduced preliminarily to Theorem 3.1, the symbol $\widetilde{\mathcal{C}}^2(K, R)$ will stand indifferently for the space

$$\begin{aligned} \mathcal{C}^{(2)}(K, \mathbf{R}) : \; = \; & \{f \in \mathcal{C}(K, \mathbf{R}) : f \text{ is the restriction to K of a twice} \\ & \text{continuously differentiable function defined on an open} \\ & \text{neighborhood of K}\} \end{aligned}$$

if $Int(K) = \emptyset$, or for the space

$$\begin{aligned} \mathcal{C}_e^2(K, \mathbf{R}) : \; = \; & \{f \in \mathcal{C}(K, \mathbf{R}) : f \text{ is twice continuously differentiable on} \\ & \text{Int(K) and its derivatives admit continuous extensions} \\ & \text{up to } K\} \end{aligned}$$

if $Int(K) \neq \emptyset$.

Moreover, by $\mathcal{F}(K, \mathbf{R}^p)$ we mean the space of all functions f from K into $\mathbf{R}^p$, while $S(K, \mathbf{R}^p)$ stands for the linear space spanned by the set $\{\psi_x : x \in K\}$.

Let E be a linear subspace of $\mathcal{F}(K, \mathbf{R})$ such that

$$\{f \in \mathcal{C}(K, \mathbf{R}) : \sup_{x \in K} \frac{|f(x)|}{1 + \|x\|^2} \in \mathbf{R}\} \subset E, \tag{3.14}$$

and denote by $pr_i : \mathbf{R}^p \mapsto \mathbf{R}$ the i-th projection on $\mathbf{R}^p$, i.e., the mapping defined by setting

$$pr_i(x) := x_i \text{ for every } x = (x_i)_{1 \le i \le p} \in \mathbf{R}^p. \tag{3.15}$$

Before we state next theorem, we remark that if $f \in \widetilde{C}^2(K, \mathbf{R})$ and each of its second order partial derivative $\frac{\partial^2 f}{\partial x_i \partial x_j}$ is uniformly continuous and bounded, then the second order derivative f'' of f is uniformly continuous and bounded on K, that is, $f \in U\widetilde{C}_b^2(K, \mathbf{R})$ according to the definition in (2.8).

Indeed, by the sake of simplicity we assume that $\mathbf{R}^p$ is endowed with the Euclidean norm $\|\cdot\|$. We recall that, if $f \in \widetilde{C}^2(K, \mathbf{R})$, then f possesses continuous first and second partial derivatives on a neighborhood of K or, respectively, on the interior of K, which are continuously extendable up to K. Moreover, if $a \in K$, then $f'(a) : \mathbf{R}^p \to \mathbf{R}$ is the linear form defined by

$$f'(a)(y) = \sum_{i=1}^{p} \frac{\partial f(a)}{\partial x_i} y_i \qquad (y = (y_i)_{1 \le i \le p} \in \mathbf{R}^p),$$

while the second derivative of f in a is the bilinear form $f''(a) : \mathbf{R}^p \times \mathbf{R}^p \to \mathbf{R}$ defined by

$$f'(a)(y, z) = \sum_{i,j=1}^{p} \frac{\partial^2 f(a)}{\partial x_i \partial x_j} y_i z_j$$

for every $y = (y_i)_{1 \le i \le p}$, $z = (z_i)_{1 \le i \le p} \in \mathbf{R}^p$.

Moreover, since

$$\|f''(a)\|_{\mathcal{L}^{(2)}(\mathbf{R}^p, \mathbf{R})} \le \sum_{i,j=1}^{p} |\frac{\partial^2 f(a)}{\partial x_i \partial x_j}|,$$

f'' is bounded on K provided that each second partial derivative of f is bounded uniformly on K; furthermore, as for any $a, b \in K$

$$\|f''(a) - f''(b)\|_{\mathcal{L}^{(2)}(\mathbf{R}^p, \mathbf{R})} \le \sum_{i,j=1}^{p} |\frac{\partial^2 f(a)}{\partial x_i \partial x_j} - \frac{\partial^2 f(b)}{\partial x_i \partial x_j}|,$$

if all second partial derivatives are uniformly continuous on K, then f'' is also uniformly continuous on K.

We are now ready to present the following result.

Theorem 3.5 *Let $\alpha_{ij}, \beta_i, \gamma, w : K \mapsto \mathbf{R}$ be given $(i, j = 1, \ldots, p)$, and consider a divergent sequence of positive integers $(\varphi(n))_{n\in\mathbf{N}}$ and a sequence $(L_n)_{n\in\mathbf{N}}$ of positive linear operators from E into $\mathcal{F}(K, \mathbf{R})$.*
Let

$$A : \widetilde{\mathcal{C}}^2(K, R) \to \mathcal{F}(K, \mathbf{R})$$

be the operator defined by setting for any $f \in \widetilde{\mathcal{C}}^2(K, R)$ and $x \in K$

$$A(f)(x) := \frac{1}{2} \sum_{i,j=1}^{p} \alpha_{ij}(x) \frac{\partial^2 f}{\partial x_i \partial x_j}(x) + \sum_{i=1}^{p} \beta_i(x) \frac{\partial f}{\partial x_i}(x) + \gamma(x) f(x). \qquad (3.16)$$

Assume that

(a) for $k = 0, 2$

$$\lim_n \|x\|^k w(x)[\varphi(n)(L_n(\mathbf{1})(x) - 1) - \gamma(x)] = 0 \qquad \text{uniformly on } K$$

(b) for $k = 0, 1$ and $i = 1, \ldots, p$

$$\lim_n \|x\|^k w(x)[\varphi(n)(L_n(pr_i \circ \psi_x)(x) - \beta_i(x)] = 0 \qquad \text{uniformly on } K;$$

(c) for every $i, j = 1, \ldots, p$

$$\lim_n w(x)[\varphi(n) L_n((pr_i \circ \psi_x)(pr_j \circ \psi_x))(x) - \alpha_{ij}(x)] = 0 \qquad \text{uniformly on } K;$$

(d)

$$\sup_{n\geq 1, x\in K} w(x)\varphi(n) L_n(d_x^2)(x) < +\infty,$$

(e) there exists some $q \in \mathbf{R}$, $q > 2$, such that $d_x^q \in E$ for all $x \in K$ and

$$\lim_n w(x)\varphi(n) L_n(d_x^q)(x) = 0$$

uniformly with respect to $x \in K$.

Then for any $f \in \widetilde{\mathcal{C}}^2(K, R)$ such that all the partial derivatives $\frac{\partial^2 f}{\partial x_i \partial x_j}$ $(1 \leq i \leq p, 1 \leq j \leq p)$ are uniformly continuous and bounded,

$$\lim_n w(x)[\varphi(n)(L_n(f) - f) - A(f)](x) = 0 \text{ uniformly on } K.$$

Furthermore, the same result holds true provided that the operators L_n are of the form (3.8) and condition (e) is replaced by the following one

(f) for every $\delta > 0$

$$\lim_n \varphi(n) w(x) \int_{K\setminus B(x,\delta)} d_x^2 \, d\mu_{n,x} = 0$$

uniformly with respect to $x \in K$.

Proof. We shall apply Theorem 3.1, part (2) and Corollary 3.2 to the present case. To this aim, consider $f \in \widetilde{\mathcal{C}}^2(K,R)$ such that $\frac{\partial^2 f}{\partial x_i \partial x_j}$ is uniformly continuous and bounded for every $i,j : 1,\dots,p$. Define for any $n \in \mathbf{N}$ the operator

$$\widetilde{L}_n : S(K,\mathbf{R}^p) \mapsto \mathcal{F}(K,\mathbf{R}^p)$$

by setting for every $\varphi \in S(K,\mathbf{R}^p)$

$$\widetilde{L}_n(\varphi) := (L_n(pr_1 \circ \varphi), L_n(pr_2 \circ \varphi), \dots, L_n(pr_p \circ \varphi)). \tag{3.17}$$

Now observe that for any linear form u on $\mathbf{R}^p$, denoted by $(u_1,\dots,u_p)$ the vector in $\mathbf{R}^p$ representing u, for every $x,y \in K$ we have

$$(u \circ \psi_x)(y) = \sum_{i=1}^{p} u_i pr_i(y-x) = \sum_{i=1}^{p} u_i (pr_i \circ \psi_x)(y);$$

therefore for any $x \in K$

$$L_n(u \circ \psi_x) = L_n(\sum_{i=1}^{p} u_i (pr_i \circ \psi_x)) = \sum_{i=1}^{p} u_i L_n(pr_i \circ \psi_x) = u \circ \widetilde{L}_n(\psi_x). \tag{3.18}$$

If we set $\beta(x) := (\beta_1(x),\dots,\beta_p(x))$ for any $x \in K$, we have for k=0,1:

$$\lim_n w(x)\|x\|^k(\varphi(n)\widetilde{L}_n(\psi_x)(x) - \beta(x)) = 0$$

uniformly on K because of hypothesis (b) and of equality (3.18), and therefore condition (ii) of Theorem 3.1 is fulfilled.
Furthermore, in order to verify that (iii) of the statement of Theorem 3.1 is satisfied by a suitable operator T, set for any $v \in \mathcal{L}^{(2)}(\mathbf{R}^p,\mathbf{R})$ and $x \in K$

$$T(v)(x) := \sum_{i,j=1}^{p} v_{ij}\alpha_{ij}(x), \tag{3.19}$$

where $(v_{ij})_{i,j=1,\dots,p}$ is the $p \times p$ matrix canonically associated to v.

In particular from (3.19) we deduce

$$T(f''(x)) = \sum_{i,j=1}^{p} \frac{\partial^2 f}{\partial x_i \partial x_j}(x)\alpha_{ij}. \qquad (x \in K)$$

We also point out that, if $\|\cdot\|_2$ denotes the Euclidean norm on $\mathbf{R}^p$, there exists $h > 0$ such that $\|\cdot\| \le h\|\cdot\|_2$. Hence, for every $v \in \mathcal{L}^{(2)}(\mathbf{R}^p,\mathbf{R})$, denoted by $(v_{ij})_{i,j=1,\dots,p}$ the relevant representing matrix, we have that

$$|\, v_{ij} \,| \le h^2 \,\|\, v \,\| \qquad \text{for every } i,j = 1,\dots,p.$$

Consider a bounded subset B of $\mathcal{L}^{(2)}(\mathbf{R}^p, \mathbf{R})$ and set $s_B := \sup_{v\in B} \|v\|$; therefore for every $x \in K$

$$
\begin{aligned}
&|w(x)(\varphi(n)L_n(v\circ\psi_x^{(2)})(x) - T(v)(x))| = \\
&\quad = |w(x)(\varphi(n)L_n\left(\sum_{i,j=1}^{p} v_{ij}(pr_i\circ\psi_x)(pr_j\circ\psi_x)\right)(x) - \sum_{i,j=1}^{p} v_{ij}\alpha_{ij}(x))| = \\
&\quad = |w(x)(\sum_{i,j=1}^{p} v_{ij}(\varphi(n)L_n((pr_i\circ\psi_x)(pr_j\circ\psi_x))(x) - \alpha_{ij}(x))| \leq \\
&\quad \leq |w(x)|h^2 s_B \sum_{i,j=1}^{p} |(\varphi(n)L_n((pr_i\circ\psi_x)(pr_j\circ\psi_x))(x) - \alpha_{ij}(x))|
\end{aligned}
$$

and so, by hypothesis (c),

$$\lim_n w(x)[\varphi(n)L_n(v\circ\psi_x^{(2)})(x) - T(v)(x)] = 0$$

uniformly on K and with respect to $v \in B$.
Therefore, Theorem 3.1, part (2) applies, and thus

$$\lim_n w(x)[\varphi(n)(L_n(f) - f) - A(f)](x) = 0$$

uniformly on K.

Finally, if each operator L_n is of the form prescribed in (3.8) and condition (f) replaces (e), by Corollary 3.2 the assertion readily follows.

□

Remarks 3.6 1). In the particular case in which K is a (possibly unbounded) real interval, the same result described above was already obtained in [2], Theorem 1.

2). Theorem 3.5 includes, as a particular case, Theorem 4.1 of [9].

4 Applications

In this Section we make use of the results of Section 3 to obtain an asymptotic formula for Bernstein-Schnabl operators defined on a bounded convex subset of a separable Banach space as well as for the multidimensional extension of the so-called Szász-Mirakjan operators.

Let X be a separable Banach space, and consider a convex and bounded subset K of X. Moreover, let $(\mu_x)_{x\in X}$ be a family of probability Borel measures on K such that

$$\int_K u_{|K} d\mu_x = u(x) \qquad \text{for any } u \in X'. \tag{4.1}$$

Set $E := \mathcal{C}_b(K, \mathbf{R})$ the space of all bounded continuous functions on K. The space E verifies (2.9) because K is bounded.

For every $n \geq 1$ consider the n-th Bernstein-Schnabl operator defined by

$$B_n(f)(x) := \int_K \cdots \int_K f\left(\frac{x_1 + \cdots + x_n}{n}\right) d\mu_x(x_1) \cdots d\mu_x(x_n) \tag{4.2}$$

($f \in E$, $x \in K$)(see [3], Chapter 6, for several additional properties of these operators).

Notice that if we consider the mappings $\pi_n : X^n \to X$ defined by

$$\pi_n(x_1, \cdots, x_n) := \frac{x_1 + \cdots + x_n}{n} \qquad (x_1, \cdots, x_n \in X) \tag{4.3}$$

and if we denote by

$$\mu_{x,n} := \pi_n(\mu_x \otimes \cdots \otimes \mu_x) \qquad (x \in X, n \geq 1) \tag{4.4}$$

the image measure of the n-times tensorial product $\mu_x \otimes \cdots \otimes \mu_x$ under π_n, the operators B_n may be put in the form (3.8) by writing

$$B_n(f)(x) = \int_K f \, d\mu_{x,n} \qquad (f \in E, n \geq 1, x \in K). \tag{4.5}$$

Moreover, by virtue of a theorem of Kolmogorov's (see [7], Corollary 9.5), for any $x \in K$ there exists a sequence $(X_{n,x})_{n\geq 1}$ of independent random variables defined on a suitable probability space $(\Omega, \mathcal{F}, P)$ and with values in K such that μ_x is the distribution of $X_{n,x}$, for all $n \geq 1$.

This argument allows us to write the operators B_n also as

$$B_n(f)(x) = \int_\Omega f\left(\frac{1}{n}\sum_{i=1}^n X_{i,x}\right) dP. \qquad (f \in E, n \geq 1, x \in K) \tag{4.6}$$

Finally we remark that for every $v \in \mathcal{L}^{(2)}(X, \mathbf{R})$ and $x \in K$

$$\int_K |v(y, y)| d\mu_x(y) < +\infty,$$

because of the boundedness of K.

We are now ready to state the following result

Theorem 4.1 *Let X be a separable Banach space, K a bounded and convex subset of X and $(\mu_x)_{x\in K}$ a family of probability Borel measures on K satisfying (4.1). Consider the sequence $(B_n)_{n\geq 1}$ of the Bernstein-Schnabl operators defined by (4.2), and a bounded function $w \in \mathcal{F}(K, \mathbf{R})$. If*

$$\sup_{n\geq 1, x\in K} \mid w(x) \mid nB_n(d_x^2)(x) < +\infty, \tag{4.7}$$

then for every $f \in U\widetilde{\mathcal{C}}_b^2(K, R)$

$$\lim_{n\to\infty} w(x)\left[n(B_n(f)(x) - f(x)) - \frac{1}{2}\left(\int_K f''(x)(y,y)d\mu_x(y) - f''(x)(x,x)\right)\right] = 0$$

uniformly with respect to $x \in K$.

Proof. We shall apply Theorem 3.1 and Corollary 3.2 to the present case, where $E := \mathcal{C}_b(K, \mathbf{R})$ and $\varphi(n) = n, (n \geq 1)$. Preliminarily observe that

$$B_n(\mathbf{1}) = \mathbf{1}, \qquad (n \geq 1)$$

and, for any $u \in X'$ and $x \in K$,

$$B_n(u \circ \psi_x) = u \circ \psi_x. \qquad (n \geq 1)$$

Moreover, for every $v \in \mathcal{L}^{(2)}(X, \mathbf{R})$ and $x \in K$

$$\begin{aligned}
& B_n(v \circ \psi_x^{(2)})(x) = \\
= \ & \int_K \cdots \int_K v\left(\frac{x_1 + \cdots + x_n}{n} - x, \frac{x_1 + \cdots + x_n}{n} - x\right) d\mu_x(x_1) \cdots d\mu_x(x_n) = \\
= \ & \int_K \cdots \int_K v\left(\frac{x_1 + \cdots + x_n}{n}, \frac{x_1 + \cdots + x_n}{n}\right) d\mu_x(x_1) \cdots d\mu_x(x_n) + \\
- \ & 2\int_K \cdots \int_K v\left(\frac{x_1 + \cdots + x_n}{n}, x\right) d\mu_x(x_1) \cdots d\mu_x(x_n) + v(x,x) = \\
= \ & \frac{1}{n^2} \sum_{i,j=1}^{n} \int_K \int_K v(x_i, x_j) d\mu_x(x_i) d\mu_x(x_j) - v(x,x) = \\
= \ & \frac{1}{n^2}\left(n\int_K v(y,y)d\mu_x(y) + n(n-1)v(x,x)\right) - v(x,x) = \\
= \ & \frac{1}{n}\left(\int_K v(y,y)d\mu_x(y) - v(x,x)\right).
\end{aligned}$$

Therefore, if we set $\gamma = \mathbf{0}$, $\beta = \mathbf{0}$, $\widetilde{B}_n = I$ for every $n \in \mathbf{N}$ (where I is the identity operator on $S(K, X)$) and if we consider the operator $T : \mathcal{L}^{(2)}(X, \mathbf{R}) \to \mathcal{F}(K, \mathbf{R})$ defined as

$$T(v)(x) := \int_K v(y,y)d\mu_x(y) - v(x,x), \qquad (v \in \mathcal{L}^{(2)}(X, \mathbf{R}), x \in K)$$

conditions (i), (ii) and (iii) of Theorem 3.1 are all satisfied.

In addition, set $D := \sup_{x\in K} \|x\|$; we shall prove that formula (3.9) is verified. Consider $\delta > 0$, $x \in K$ and $n \in \mathbf{N}$, and take (4.5) and (4.6) into account; then

$$\begin{aligned} \int_{K\setminus B(x,\delta)} d_x^2 \, d\mu_{x,n} &\leq 4D^2 \int_K \mathbf{1}_{K\setminus B(x,\delta)} d\mu_{x,n} = \\ &= 4D^2 \int_\Omega \mathbf{1}_{K\setminus B(x,\delta)} \circ \left(\frac{1}{n}\sum_{i=1}^n X_{i,x}\right) dP = \\ &= 4D^2 P\{\| \frac{1}{n}\sum_{i=1}^n X_{x,i} - x \| > \delta\}. \end{aligned} \tag{1}$$

By the strong law of large numbers there is some $\nu \in \mathbf{N}$ such that for any $n \in \mathbf{N}$, $n \geq \nu$

$$E(\| \frac{1}{n}\sum_{i=1}^n X_{x,i} - x \|) < \frac{\delta}{2};$$

therefore, following the proof of a Lemma by Beśka and Dziedziul (in [8], pg. 715), for any such n we obtain

$$\begin{aligned} P\{\| \frac{1}{n}\sum_{i=1}^n X_{x,i} - x \| > \delta\} &\leq P\{\| \frac{1}{n}\sum_{i=1}^n X_{x,i} - x \| > \frac{\delta}{2} + E \| \frac{1}{n}\sum_{i=1}^n X_{x,i} - x \|\} \\ &\leq P\{\| \sum_{i=1}^n X_{x,i} - nx \| - E(\| \sum_{i=1}^n X_{x,i} - nx \|) > \frac{n\delta}{2}\}. \end{aligned}$$

By applying an inequality established in ([11], formula (2.19), page 184) to the independent sequence of bounded random variables on K

$$X_n := X_{x,n} - x, \qquad (n \in \mathbf{N})$$

we deduce that

$$\begin{aligned} &P\{\|\sum_{i=1}^n X_{x,i} - nx \| - E(\| \sum_{i=1}^n X_{x,i} - nx \|) > \frac{n\delta}{2}\} \leq \\ &\leq 2exp(-\frac{n^2\delta^2}{32\sum_{i=1}^n \|X_n\|^2}) \leq 2exp(-\frac{n^2\delta^2}{32n4D^2}) = 2exp(-\frac{n\delta^2}{128D^2}); \end{aligned}$$

combining this last inequality with (1) we obtain that for any $\delta > 0$ and $n \geq \nu$

$$\int_{K\setminus B(x,\delta)} d_x^2 \, d\mu_{x,n} \leq 8D^2 exp(-\frac{n\delta^2}{128D^2})$$

uniformly with respect to $x \in K$, so that condition (3.9) of Corollary 3.2 is also fulfilled. The assertion follows straightforwardly from this last mentioned corollary.

□

Remarks 4.2 1) If $(X, <\cdot,\cdot>)$ is a Hilbert space, condition (4.7) holds true with $w = \mathbf{1}$; indeed, for any $n \geq 1$ and $x \in K$

$$\begin{aligned} & B_n(d_x^2)(x) = \\ = & \int_K \cdots \int_K < \frac{x_1 + \cdots + x_n}{n} - x, \frac{x_1 + \cdots + x_n}{n} - x > d\mu_x(x_1) \cdots d\mu_x(x_n) \\ = & \frac{1}{n}\left(\int_K \|y\|^2 d\mu_x(y) - \|x\|^2\right). \end{aligned}$$

2) Theorem 4.1 was also obtained by Beśka and Dziedziul in [8], Theorem 4.2, in the special case in which K is a simplex and $w = \mathbf{1}$.

If K is a convex and compact subset of some Euclidean space, a similar result was previously proved in [1], Proposition 2.1 (see also [3], Theorem 6.2.5).

As a second application of the results of the previous section we examine an asymptotic formula for the sequence of the p-dimensional Szász-Mirakjan operators, $p \geq 1$, which has been introduced by a more general scheme by Nishishirao in [10], Section 4.

Consider the space $\mathbf{R}^p$ endowed with an arbitrary norm $\|\cdot\|$ and set $K := \{x = (x_i)_{1 \leq i \leq p} \in \mathbf{R}^p \mid x_i \geq 0 \text{ for every } i = 1, \ldots, p\}$ and

$$\begin{aligned} E := \{f \in \mathcal{C}(K, \mathbf{R}) \quad : \quad & |f(x)| \leq M\|x\|^m \text{ for any } x \in K \\ & \text{and for some } M > 0 \text{ and } m \in \mathbf{N}\}. \end{aligned}$$

Next we introduce the n-th p-dimensional Szász-Mirakjan operator, defined for any $f \in E$ and $(x_1, \ldots, x_p) \in K$ as

$$A_{n,p}(f)(x_1, \ldots, x_p) := e^{-n\sum_{i=1}^p x_i} \sum_{k_1=0}^{\infty} \cdots \sum_{k_p=0}^{\infty} f\left(\frac{k_1}{n}, \ldots, \frac{k_p}{n}\right) \prod_{i=1}^{p} \frac{(nx_i)^{k_i}}{k_i!}. \tag{4.3}$$

The positive operators $A_{n,p}$ are well defined on E. Assume, indeed, that a function $f \in \mathcal{C}(K, \mathbf{R})$ satisfies a growth condition of the form $|f(x)| \leq M\|x\|^m$ $(x \in K)$ for some $m \in \mathbf{N}$ and $M \geq 0$. If $m = 0$, f is bounded and clearly the series in (4.3) are convergent; if $m \geq 1$, since the norm $\|\cdot\|$ is equivalent to the l_m-norm

$$\|(x_1, \ldots, x_p)\|_m := \left(\sum_{i=1}^{p} x_i^m\right)^{1/m} \qquad ((x_1, \ldots, x_p) \in K),$$

there exists a further constant $R \geq 0$ such that

$$|f(x)| \leq R\|x\|_m^m \qquad (x \in K).$$

Therefore for every $x = (x_1, \ldots, x_p) \in K$ the series in (4.3) is convergent as a linear combination of the sum of the series

$$\sum_{k_i=0}^{\infty} \left(\frac{k_i}{n}\right)^m \frac{(nx_i)^{k_i}}{k_i!} \qquad (i = 1, \ldots, p)$$

which are convergent by the ratio test.

Among other things, the operator $A_{n,p}$ may itself be put in the form (3.8); indeed, consider the following probability measure on $\mathbf{R}$:

$$\pi_{n,y} := \sum_{h=0}^{\infty} e^{-ny} \frac{(ny)^h}{h!} \varepsilon_{\frac{h}{n}} \qquad (y \geq 0, n \geq 1),$$

where $\varepsilon_{\frac{h}{n}}$ denotes the unit mass at $\frac{h}{n}$ and set, for any $x := (x_1, \ldots, x_p) \in K$

$$\tau_{n,x} := \bigotimes_{i=1}^{p} \pi_{n,x_i}; \qquad (n \geq 1)$$

then we may write

$$A_{n,p}(f)(x) = \int_K f d\tau_{n,x}. \qquad (f \in E, x \in K, n \geq 1)$$

Let us now state the pre-announced asymptotic formula regarding the sequence $(A_{n,p})_{n\geq 1}$.

Theorem 4.3 *For any $f \in \widetilde{\mathcal{C}}^2(K, R)$ such that every partial derivative $\frac{\partial^2 f}{\partial x_i \partial x_j}$ $(1 \leq i, j \leq p)$ is uniformly continuous and bounded,*

$$\lim_n \frac{1}{1 + \|x\|^2} \left[n \left(A_{n,p}(f) - f \right)(x) - \frac{1}{2} \sum_{i=1}^{p} x_i \frac{\partial^2 f}{\partial x_i^2}(x) \right] = 0 \qquad (4.4)$$

uniformly with respect to $x = (x_1, \ldots, x_p) \in K$.

In particular, for every compact subset Q of K

$$\lim_n n \left(A_{n,p}(f) - f \right)(x) = \frac{1}{2} \sum_{i=1}^{p} x_i \frac{\partial^2 f}{\partial x_i^2}(x)$$

uniformly with respect to $x = (x_i)_{1\leq i\leq p} \in Q$.

Before we prove (4.4), we remark that it has already been established in the one-dimensional case for two different classes of operators which both include the Szász-Mirakjan ones (see Theorem 2.3 in [4] and Proposition 3.1 in [5]).

Let us now proceed to verify that conditions (a)-(e) of Theorem 3.5 hold true.

To this purpose set $\gamma = \mathbf{0}$, $\beta_i = \mathbf{0}$, $\alpha_{i,i} = pr_i$, and $\alpha_{i,j} = \mathbf{0}$ if $i \neq j$, for every $i, j = 1, \ldots, p$, and moreover set $w(x) := \frac{1}{1+\|x\|^2}$ $(x \in K)$ and $\varphi(n) := n$ for any $n \in \mathbf{N}$.

Notice that, with this coefficients, the differential operator (3.16) prescribed in the statement of Theorem 3.5 has the form

$$A(f)(x_1, \ldots, x_p) = \frac{1}{2} \sum_{i=1}^{p} x_i \frac{\partial^2 f}{\partial x_i^2}(x_1, \ldots, x_p)$$

for all $f \in \widetilde{\mathcal{C}}^2(K, R)$ and $(x_1, \ldots, x_p) \in K$.

For any $n \geq 1$, $i, j = 1, \ldots, p$ and $x = (x_1, \ldots, x_p) \in K$, we get

$$A_{n,p}(\mathbf{1}) = \mathbf{1},$$

$$A_{n,p}(pr_i) = pr_i,$$

$$A_{n,p}(pr_i^2) = pr_i^2 + \frac{1}{n} pr_i,$$

and hence

$$A_{n,p}(pr_i \circ \psi_x)(x) = 0,$$

$$A_{n,p}((pr_i \circ \psi_x)(pr_j \circ \psi_x))(x) = 0 \qquad if\ i \neq j,$$

$$A_{n,p}\left((pr_i \circ \psi_x)^2\right)(x) = \frac{pr_i}{n}(x),$$

so that (a), (b) and (c) are satisfied; moreover, as

$$A_{n,p}(d_x^2)(x) = \frac{1}{n} \sum_{i=1}^{p} x_i,$$

we readily deduce that

$$\sup_{x \in K, n \geq 1} \frac{n A_{n,p}(d_x^2)(x)}{1 + \|x\|^2} < +\infty,$$

and thus also formula (d) holds true.

Next fix $x = (x_1, \ldots, x_p) \in K$ and observe that for any $y \in K$, $y = (y_1, \ldots, y_p)$,

$$d_x^4(y) = \left(\sum_{i=1}^{p} (x_i - y_i)^2 \right)^2 \leq p^2 \sum_{i=1}^{p} (x_i - y_i)^4,$$

that is

$$d_x^4 \leq p^2 \sum_{i=1}^{p} (pr_i \circ \psi_x)^4,$$

and thus for any $n \geq 1$

$$A_{n,p}(d_x^4)(x) \leq p^2 \sum_{i=1}^{p} A_{n,p}((pr_i \circ \psi_x)^4)(x).$$

On the other hand, for every $i = 1, \ldots, p$ it can be proved that

$$A_{n,p}((pr_i \circ \psi_x)^4)(x) = \sum_{k=0}^{\infty} e^{-nx_i} (\frac{k}{n} - x_i)^4 \frac{(nx_i)^k}{k!} = \frac{3}{n^2} x_i^2 + \frac{x_i}{n^3}$$

(this equality is fully detailed in [12], Proposition 4; otherwise one may see it, in a more general case, in the proof of Proposition 3.1 of [5]).

Summing up we obtain for every $x = (x_1, \ldots, x_p) \in K$ and $n \geq 1$

$$A_{n,p}(d_x^4)(x) \leq p^2 \sum_{i=1}^{p} (\frac{3}{n^2} x_i^2 + \frac{x_i}{n^3}),$$

which implies that

$$\lim_n \frac{n A_{n,p}(d_x^4)(x)}{1 + \|x\|^2} = 0 \quad \text{uniformly on K}$$

and thus, finally, also condition (e) of Theorem 3.5 is satisfied with q=4.

References

[1] F. Altomare, *Lototsky-Schnabl operators on compact convex sets and their associated limit semigroups*, Monatsh. Math., 114, (1992), 1-13.

[2] F. Altomare and R. Amiar, *Asymptotic formulae for positive linear operators*, Math. Balkanica, **16** (2002), 283-304.

[3] F. Altomare and M. Campiti, Korovkin-type Approximation Theory and its Applications, de Gruyter Studies in Mathematics, Vol 17, Walter de Gruyter and Co., Berlin, New York, 1994.

[4] F. Altomare and I. Carbone, *On some degenerate differential operators on weighted function Spaces*, J. Math. Anal. Appl., **213**, (1997), 308-333.

[5] F. Altomare and I. Raşa, *On a class of exponential-type operators and their limit semigroups*, preprint, 2004.

[6] A. Avez, Differential calculus, J. Wiley and Sons Ltd., Chichester, 1986.

[7] H. Bauer, Probability Theory, de Gruyter Studies in Mathematics, Vol. 23, Walter de Gruyter and Co., Berlin, 1996.

[8] M. Beśka and K. Dziedziul, *Saturation theorems for interpolation and the Bernstein- Schnabl operators*, Math. Comp., Vol **70**, N. 234 (2000), 167-241.

[9] W. Dahmen and C. A. Micchelli, *Convexity and Bernstein Polynomials on k-symploids*, Acta Math. Appl. Sinica, **6**, No. 1, (1990), 50-66.

[10] T. Nishishirao, *Refinements of Korovkin-type Approximation processes*, Rend. Circ. Mat. Palermo, (2) Suppl. **68**, Vol. II, (2002), 711-725.

[11] G. Pisier, *Probabilistic methods in the geometry of Banach spaces*, in Lecture Notes in Math., Vol **1206**, Springer-Verlag, (1986), 167-241.

[12] Y. Suzuki, *Saturation of local approximation by linear positive operators of Bernstein type*, Tohoku Math. J., Vol **19**, N. 4 (1967), 429-453.

JOURNAL OF CONCRETE AND APPLICABLE MATHEMATICS,VOL.3,NO.4,437-455,
2005,COPYRIGHT 2005 EUDOXUS PRESS,LLC

Generalised Weighted Trapezoidal Rules and Relationship to Ostrowski Results

P. Cerone and J. Roumeliotis
School of Computer Science and Mathematics
Victoria University of Technology
PO Box 14428, Melbourne City MC
Victoria 8001, Australia.
Email: pc@matilda.vu.edu.au john.roumeliotis@vu.edu.au

Abstract

The generalised weighted trapezoidal rule is investigated which involves $f^{(n)}(t)$ of bounded variation. If the weight is taken to be identically unity, then previous results are recaptured. A comparison with weighted Ostrowski results is made and it is demonstrated that if the weight is symmetric about the midpoint over the interval $[a,x)$ and $(x,b]$ then the bounds are the same. In particular, if the weight is unity, then the generalised trapezoidal and Ostrowski results produce the same bounds.

Keywords: Weighted Rules, Generalised Trapezoidal Bounds, Ostrowski.
AMS Subject Classification Codes: Primary 26D15; Secondary 65D30.

1 Introduction

K.S.K. Iyengar [11], by means of geometrical consideration, has proved the following theorem.

Theorem 1 *Let f be a differentiable function on $[a,b]$ and $|f'(x)| \leq M$. Then*

$$\left| \int_a^b f(x)\,dx - \frac{1}{2}(b-a)(f(a)+f(b)) \right| \leq \frac{M(b-a)^2}{4} - \frac{1}{4M}(f(b)-f(a))^2. \tag{1}$$

(See also [18, p. 471 - 474] for related results). Further generalisations were also given by Agarwal and Dragomir [13], and Cerone and Dragomir [15].

In [19], the following generalisation of Theorem 1 is proved analytically.

Theorem 2 *Let $f(x)$ be a differentiable function defined on $[a,b]$ and $|f'(x)| \leq M$ for every $x \in (a,b)$. If $p(x)$ is an integrable function on (a,b) such that*

$$0 < c \leq p(x) \leq \lambda c \quad (\lambda \geq 1,\ x \in [a,b]),$$

then the following inequality holds

$$\left| A(f;p) - \frac{1}{2}(f(a)+f(b)) \right| \leq \frac{M(b-a)}{2} \cdot \frac{(\lambda+q)\left(1-q^2\right)+2(\lambda-1)q}{2\lambda(1+q)-(\lambda-1)\left(1+q^2\right)}, \tag{2}$$

where A and q are defined by

$$A(f;p) = \frac{\int_a^b p(x) f(x)\,dx}{\int_a^b p(x)\,dx} \quad \text{and} \quad q = \frac{|f(b)-f(a)|}{M(b-a)}.$$

Cerone and Dragomir [6] also proved the following weighted trapezoidal result.

Theorem 3 *Let $f : I \subseteq \mathbb{R} \to \mathbb{R}$ be a differentiable mapping on $\mathring{I}$ (the interior of I) and $[a,b] \subset \mathring{I}$ with $M = \sup_{x\in[a,b]} f'(x) < \infty$, $m = \inf_{x\in[a,b]} f'(x) > -\infty$ be the first moment of $w(\cdot)$ on $[a,b]$. If f' is integrable on $[a,b]$, then the following inequality*

$$\left| \int_a^b w(x) f(x)\,dx - \frac{\nu}{2}[f(a)+f(b)] - m\left(\frac{a+b}{2}\right)[b-a-\nu] \right| \leq \frac{\nu}{2}(b-a)(S-m) \leq \frac{M-m}{2}\nu(b-a), \tag{3}$$

where S is the slope of the secant on $[a,b]$.

The well-known Ostrowski inequality is given by the following theorem [19],

Theorem 4 *Let f be a differentiable function on $[a,b]$ and let $|f'(x)| \leq M$ on $[a,b]$. Then, for every $x \in [a,b]$,*

$$\left| f(x) - \frac{1}{b-a}\int_a^b f(t)\,dt \right| \leq \left[\frac{1}{4} + \frac{\left(x - \frac{a+b}{2}\right)^2}{(b-a)^2} \right](b-a)M. \tag{4}$$

A weighted version of the Ostrowski inequality (4) for Hölder mappings has been given in [8] by Dragomir et al. Other results related to the Ostrowski inequality may be viewed in [2], [3], [20], [23] and a book devoted to Ostrowski type results edited by Dragomir and Rassias [9]. A weighted multidimensional generalisation of Ostrowski's inequality was treated by Milovanović [14].

In a recent thorough article Matić, Pečarić and Ujević [13] obtained weighted $n-$time differentiable Ostrowski type results in terms of a variety of norms.

The norms, including when $f^{(n)}$ is of bounded variation, Hölder continuous and differentiable allowing for $f^{(n+1)} \in L_p[a,b]$ in terms of the Lebesgue norms $\|\cdot\|_p$, $p \geq 1$.

It is the intention of the current paper to examine bounds for the generalised weighted trapezoidal functional. Bounds will be provided assuming $f^{(n)}(\cdot)$ to be of bounded variation, absolutely continuous, Lipschitzian and monotonic. Placing restriction on the weight function provides more explicit coarser bounds. This is accomplished through the use of a result due to Karamata [12]. Qi [21] and [22] examines weighted trapezoidal bounds using a Taylor series argument.

Following the presentation of some notation, identities are obtained for our functional of interest in Section 2. Various bounds are developed in Section 3 while in Section 4, the relationship between the trapezoidal and corresponding Ostrowski functionals is investigated. It is demonstrated that the bounds are the same if the weight function is symmetric over the mid-points of the respective intervals $[a,x)$ and $(x,b]$.

2 Some Notation and an Identity

Before proceeding to develop an identity, it is worthwhile to introduce some notation.

Let $w(\cdot)$ be a weight function and suppose that $w : [a,b] \to (0,\infty)$ is integrable on the interval $[a,b]$ and such that

$$0 < \int_a^b w(t)\,dt < \infty.$$

Also, let

$$m_k(c,d;w) = \int_c^d u^k w(u)\,du \tag{5}$$

represent the k^{th} moment about the origin of the weight function $w(\cdot)$ over the interval $[c,d] \subseteq [a,b]$. Further, let

$$\begin{aligned} 0 \leq M_n(a,x;w) &= \frac{1}{n!}\int_a^x (u-a)^n w(u)\,du \\ &= \frac{1}{n!}\sum_{k=0}^{n} \left(\begin{array}{c} n \\ k \end{array} \right) (-a)^{n-k} m_k(a,x;w) \end{aligned} \tag{6}$$

and

$$\begin{aligned} 0 &\leq M_n(x,b;w) = \frac{1}{n!}\int_x^b (b-u)^n w(u)\,du \\ &= \frac{1}{n!}\sum_{k=0}^{n} \left(\begin{array}{c} n \\ k \end{array} \right) b^{n-k}(-1)^k m_k(x,b;w). \end{aligned} \tag{7}$$

It may be observed that for $x \in [a,b]$

$$M_0(a,b;w) = M_0(a,x;w) + M_0(x,b;w) = \int_a^b w(t)\,dt = m_0(a,b;w)$$

and

$$M_n(a,x;1)=\frac{(x-a)^{n+1}}{(n+1)!},\quad M_n(x,b;1)=\frac{(b-x)^{n+1}}{(n+1)!}. \tag{8}$$

We introduce the kernel

$$Q_n(x,t;w):=\begin{cases}\dfrac{1}{(n-1)!}\displaystyle\int_x^t (t-u)^{n-1}w(u)\,du, & n\in\mathbb{N},\\ w(t), & n=0,\end{cases}\qquad x,t\in[a,b] \tag{9}$$

which satisfies

$$\frac{\partial Q_n}{\partial t}=Q_{n-1},\quad n\in\mathbb{N}. \tag{10}$$

The kernel may further be written, using (6) and (7), as

$$Q_n(x,t;w):=\begin{cases}(-1)^n M_{n-1}(t,x;w), & a\le t\le x,\\ M_{n-1}(x,t;w), & x<t\le b,\end{cases}\qquad n\in\mathbb{N} \tag{11}$$

and $Q_0(x,t;w)=w(t)$.

Further, define the functional

$$\begin{aligned}T_n(a,x,b;f;w)&=\int_a^b w(t)f(t)\,dt\\ &\quad-\sum_{k=0}^{n}\left[M_k(a,x;w)f^{(k)}(a)+(-1)^k M_k(x,b;w)f^{(k)}(b)\right]\end{aligned} \tag{12}$$

for $f:[a,b]\to\mathbb{R}$, $x\in[a,b]$ and $w(\cdot)$ is a weight function with $M_k(\cdot,\cdot;w)$ as defined by (6) and (7). The following theorem holds.

Theorem 5 *Let $f:[a,b]\to\mathbb{R}$ with $a<b$. For $n=0,1,2,\ldots$ let $Q_{n+1}(x,t;w)$ be as given by (9). Further, suppose that for some $n\in\mathbb{N}\cup\{0\}$, $f^{(n)}(t)$ exists for $t\in[a,b]$, where $f^{(0)}(t)\equiv f(t)$ then for $f^{(n)}(\cdot)$ of bounded variation the identity*

$$T_n(a,x,b;f;w)=(-1)^{n+1}\int_a^b Q_{n+1}(x,t;w)\,df^{(n)}(t) \tag{13}$$

holds where T and Q_{n+1} are as defined by (10) and (9) respectively.

Proof. Before proceeding with the proof it is worthwhile to firstly note that the solution to the recurrence relation

$$u_n=a_n-u_{n-1}\quad\text{for}\quad n=1,2,\ldots \tag{14}$$

is given explicitly as

$$u_n=\sum_{k=1}^{n}(-1)^{n-k}a_k+(-1)^n u_0. \tag{15}$$

Now, let

$$I_n = \int_a^b Q_{n+1}(x,t;w)\, df^{(n)}(t) \tag{16}$$

then integration by parts of the Riemann-Stieltjes integral gives

$$I_n = Q_{n+1}(x,t;w)\, f^{(n)}(t)\Big]_a^b + \int_a^b \frac{\partial Q_{n+1}}{\partial t} f^{(n)}(t)\, dt \tag{17}$$

and so

$$I_n = A_n(a,x,b;w;f) - I_{n-1}, \quad n = 1,2,\ldots, \tag{18}$$

where, to obtain (18) from (17), we have used

$$A_n(a,x,b;w;f) = Q_{n+1}(x,b;w)\, f^{(n)}(b) - Q_{n+1}(x,a;w)\, f^{(n)}(a) \tag{19}$$

and (10) and the fact that for $f^{(n-1)}(t)$ differentiable, $df^{(n-1)}(t) = f^{(n)}(t)\, dt$.

To obtain I_0 we may either use integration by parts from (16) or equivalently extend the validity of (18) to $n = 0$, producing

$$I_0 = A_0(a,x,b;w;f) - I_{-1},$$

where from (16)

$$I_{-1} = \int_a^b Q_0(x,t;w)\, f^{(0)}(t)\, dt$$

and so from (9)

$$I_{-1} = \int_a^b w(t)\, f(t)\, dt.$$

That is,

$$I_0 = A_0(a,x,b;w;f) - \int_a^b w(t)\, f(t)\, dt. \tag{20}$$

The solution of (18) on comparison with (14) and (15) upon using (20) is given by

$$I_n = (-1)^{n+1} \int_a^b w(t)\, f(t)\, dt + \sum_{k=0}^{n} (-1)^{n-k} A_k(a,x,b;w;f).$$

Thus,

$$(-1)^{n+1} I_n = \int_a^b w(t)\, f(t)\, dt - \sum_{k=0}^{n} (-1)^k A_k(a,x,b;w;f) \tag{21}$$

where from (19) and (11)

$$A_k(a,x,b;w;f) = M_k(x,b;w)\, f^{(k)}(b) - (-1)^{k+1} M_k(a,x;w)\, f^{(k)}(a)$$

and so from (21) the identity (13) is procured. ■

Remark 1 *If $f^{(n)}(t)$ is absolutely continuous on $[a,b]$, then it is differentiable and $df^{(n)}(t) = f^{(n+1)}(t)\,dt$ giving from (13) the identity*

$$T_n(a,x,b;f;w) = (-1)^{n+1}\int_a^b Q_{n+1}(x,t;w)\,f^{(n+1)}(t)\,dt, \tag{22}$$

where T_n and Q_n are given by (12) and (9) or (11) respectively.

Remark 2 *It $w(t) \equiv 1$ then from (9)*

$$Q_n(x,t;1) = \frac{(t-x)^n}{n!}$$

and from (6) and (7)

$$M_k(a,x;1) = \frac{(x-a)^{k+1}}{(k+1)!} \quad \text{and} \quad M_k(x,b;1) = \frac{(b-x)^{k+1}}{(k+1)!}.$$

Further, from (12) and (22)

$$\begin{aligned} T_n(a,x,b;f;1) &= \int_a^b f(t)\,dt - \sum_{k=0}^{n}\frac{(x-a)^{k+1}f^{(k)}(a) + (-1)^k(b-x)^{k+1}f^{(k)}(b)}{(k+1)!} \\ &= (-1)^{n+1}\int_a^b \frac{(t-x)^{n+1}}{(n+1)!}f^{(n+1)}(t)\,dt \end{aligned}$$

is the identity obtained in Cerone et al. [7], giving the non-weighted $n-$time differentiable generalised trapezoidal identity.

3 Inequalities for the Generalised Weighted Trapezoidal Rule

The following well known lemmas will prove useful for procuring bounds for a Riemann-Stieltjes integral. They will be stated here for lucidity.

Lemma 1 *Let $g, v : [a,b] \to \mathbb{R}$ be such that g is continuous and v is of bounded variation on $[a,b]$. Then the Riemann-Stieltjes integral $\int_a^b g(t)\,dv(t)$ exists and is such that*

$$\left|\int_a^b g(t)\,dv(t)\right| \leq \sup_{t\in[a,b]}|g(t)|\bigvee_a^b(v), \tag{23}$$

where $\bigvee_a^b(v)$ is the total variation of v on $[a,b]$.

Lemma 2 *Let $g, v : [a,b] \to \mathbb{R}$ be such that g is Riemann integrable on $[a,b]$ and v is $L-$Lipschitzian on $[a,b]$. Then*

$$\left|\int_a^b g(t)\,dv(t)\right| \leq L\int_a^b |g(t)|\,dt \tag{24}$$

with v is $L-$Lipschitzian if it satisfies

$$|v(x)-v(y)| \leq L|x-y|$$

for all $x, y \in [a,b]$.

Lemma 3 *Let $g, v : [a,b] \to \mathbb{R}$ be such that g is Riemann integrable on $[a,b]$ and v is monotonic nondecreasing on $[a,b]$. Then*

$$\left|\int_a^b g(t)\,dv(t)\right| \leq \int_a^b |g(t)|\,dv(t). \tag{25}$$

It should be noted that if v is nonincreasing, then $-v$ is nondecreasing.

Theorem 6 *Let the conditions of Theorem 5 continue to hold so that $f^{(n)}(t)$ is of bounded variation for $t \in [a,b]$. Then we have for all $x \in [a,b]$ and $n \in \mathbb{N} \cup \{0\}$*

$$\begin{aligned} &|T_n(a,x,b;f;w)| \\ &= \begin{cases} [M_n(a,x;w) + M_n(x,b;w) \\ \quad + |M_n(a,x;w) - M_n(x,b;w)|] \times \frac{1}{2}\bigvee_a^b \left(f^{(n)}\right), \\ \\ [M_{n+1}(a,x;w) + M_{n+1}(x,b;w)]\,L, \qquad f^{(n)}\ L-Lipschitzian \\ \\ M_n(a,x;w)\left[f^{(n)}(x) - f^{(n)}(a)\right] \\ \quad + M_n(x,b;w)\left[f^{(n)}(b) - f^{(n)}(x)\right], \quad f^{(n)} \quad monoton.\ nondecr. \end{cases} \end{aligned} \tag{26}$$

where $T_n(a,x,b;f;w)$ is given by (12) and $M_n(a,x;w)$, $M_n(x,b;w)$ by (6) and (7).

Here, by $\bigvee_a^b(h)$ we signify the total variation of $h(t)$ for $t \in [a,b]$. That is $\bigvee_a^b(h) = \int_a^b |dh(t)|$.

Proof. Taking the modulus of identity (13) and using Lemma 1, we have

$$\begin{aligned} |T_n(a,x,b;f;w)| &= \left|\int_a^b Q_{n+1}(x,t;w)\,df^{(n)}(t)\right| \\ &\leq \sup_{t\in[a,b]} |Q_{n+1}(x,t;w)| \bigvee_a^b \left(f^{(n)}\right). \end{aligned} \tag{27}$$

Now, from (9)

$$\begin{aligned} &\sup_{t\in[a,b]} |Q_{n+1}(x,t;w)| \\ &\quad = \frac{1}{n!}\max\left\{\int_a^x (u-a)^n w(u)\,du, \int_x^b (b-u)^n w(u)\,du\right\} \\ &\quad = \max\{M_n(a,x;w), M_n(x,b;w)\} \end{aligned} \tag{28}$$

and so using the fact that $\max\{X,Y\} = \frac{1}{2}[X+Y+|X-Y|]$ gives the first inequality in (26) upon utilising (27).

For $f^{(n)}(\cdot)$ $L-$Lipschitzian on $[a,b]$, then from Lemma 2 and (27)

$$\begin{aligned} |T_n(a,x,b;f;w)| &= \left| \int_a^b Q_{n+1}(x,t;w)\, df^{(n)}(t) \right| \\ &\leq L \int_a^b |Q_{n+1}(x,t;w)|\, dt. \end{aligned} \tag{29}$$

Using the definition (9), we may notice that

$$Q_{n+1}(x,t;w) = \begin{cases} \dfrac{(-1)^{n+1}}{n!} \displaystyle\int_t^x (u-t)^n w(u)\, du, & t \in [a,x] \\ \dfrac{1}{n!} \displaystyle\int_x^t (t-u)^n w(u)\, du, & t \in (x,b] \end{cases} \tag{30}$$

and so

$$\begin{aligned} & n! \int_a^b |Q_{n+1}(x,t;w)|\, dt \\ & \qquad = \int_a^x \int_t^x (u-t)^n w(u)\, dudt + \int_x^b \int_x^t (t-u)^n w(u)\, dudt. \end{aligned} \tag{31}$$

We may simplify the expression on the right by an interchange of the order of integration to give

$$\begin{aligned} \int_a^x \int_t^x (u-t)^n w(u)\, dudt &= \int_a^x w(u) \int_a^u (u-t)^n\, dtdu \\ &= \tfrac{1}{n+1} \int_a^x (u-a)^{n+1} w(u)\, du \end{aligned} \tag{32}$$

and in a similar fashion

$$\int_x^b \int_x^t (t-u)^n w(u)\, dudt = \frac{1}{n+1} \int_x^b (b-u)^{n+1} w(u)\, du$$

Hence, from (31),

$$\int_a^b |Q_{n+1}(x,t;w)|\, dt = M_{n+1}(a,x;w) + M_{n+1}(x,b;w) \tag{33}$$

giving the second inequality in (26) upon utilising (29).

For the final inequality in (26), when $f^{(n)}(t)$ is monotonic nondecreasing on $[a,b]$, we use Lemma 3 and thus, from identity (13)

$$\begin{aligned} |T_n(a,x,b;f;w)| &= \left| \int_a^b Q_{n+1}(x,t;w)\, df^{(n)}(t) \right| \\ &\leq \int_a^b |Q_{n+1}(x,t;w)|\, df^{(n-1)}(t). \end{aligned} \tag{34}$$

Two cases need to be treated. For $n=0$ then

$$|T_0(a,x,b;f;w)| \leq \int_a^b |Q_1(x,t;w)|\, df(t),$$

where, from (9),

$$|Q_1(x,t;w)| = \begin{cases} \displaystyle\int_t^x w(u)\, du, & t \in [a,x], \\ \displaystyle\int_x^t w(u)\, du, & t \in (x,b]. \end{cases}$$

Thus,

$$\begin{aligned} &|T_0(a,x,b;f;w)| \\ &\quad \leq \int_a^x \left(\int_t^x w(u)\, du \right) df(t) + \int_x^b \left(\int_x^t w(u)\, du \right) df(t) \\ &\quad = \left(\int_t^x w(u)\, du \right) f(t) \Big]_{t=a}^{x} + \int_a^x w(t) f(t)\, dt \\ &\qquad + \left(\int_x^t w(u)\, du \right) f(t) \Big]_{t=x}^{b} - \int_x^b w(t) f(t)\, dt \\ &\quad = -M_0(a,x;w) f(a) \\ &\qquad + \int_a^x w(t) f(t)\, dt + M_0(x,b;w) f(b) - \int_x^b w(t) f(t)\, dt \\ &\quad \leq M_0(a,x;w)[f(x) - f(a)] + M_0(x,b;w)[f(b) - f(x)]. \end{aligned} \tag{35}$$

Here we have used the fact that if $g(t) > 0$ and $f(t)$ monotonic nondecreasing for $t \in [a,b]$, then

$$\begin{aligned} \int_a^b g(t) f(t)\, dt &\leq f(b) \int_a^b g(t)\, dt \quad \text{and} \\ -\int_a^b g(t) f(t)\, dt &\leq -f(a) \int_a^b g(t)\, dt. \end{aligned} \tag{36}$$

For $n \in \mathbb{N}$ then

$$\begin{aligned} & n! \int_a^b |Q_{n+1}(x,t;w)| \, df^{(n)}(t) \\ & \quad = \int_a^x \left(\int_t^x (u-t)^n w(u) \, du \right) df^{(n)}(t) \\ & \qquad + \int_x^b \left(\int_x^t (t-u)^n w(u) \, du \right) df^{(n)}(t) \\ & \quad := A_n + B_n. \end{aligned} \tag{37}$$

Integration by parts produces on using the Leibnitz rule,

$$\begin{aligned} A_n &= \left(\int_t^x (u-t)^n w(u) \, du \right) f^{(n)}(t) \Bigg]_{t=a}^{x} \\ &\quad + n \int_a^x \left(\int_t^x (u-t)^{n-1} w(u) \, du \right) f^{(n)}(t) \, dt \\ &= -n! M_n(a,x;w) f^{(n)}(a) + n \int_a^x \left(\int_t^x (u-t)^{n-1} w(u) \, du \right) f^{(n)}(t) \, dt \\ &\leq n! M_n(a,x;w) \left[f^{(n)}(x) - f^{(n)}(a) \right] \end{aligned}$$

and

$$\begin{aligned} B_n &= \left(\int_x^t (t-u)^n w(u) \, du \right) f^{(n)}(t) \Bigg]_{t=x}^{b} \\ &\quad - n \int_x^b \left(\int_x^t (t-u)^{n-1} w(u) \, du \right) f^{(n)}(t) \, dt \\ &= n! M_n(x,b;w) f^{(n)}(b) - n \int_x^b \left(\int_x^t (t-u)^{n-1} w(u) \, du \right) f^{(n)}(t) \, dt \\ &\leq n! M_n(x,b;w) \left[f^{(n)}(b) - f^{(n)}(x) \right], \end{aligned}$$

where we have used the monotonicity of $f^{(n)}(\cdot)$ via (36) together with (31) and (33) to obtain the upper bounds.

Substituting A_n and B_n into (37) and (34) and further recognising that it subsumes the result for $n = 0$ as given by (35), then the last inequality in (26) results. ■

Remark 3 *For the monotonic nondecreasing result in (26) a tighter bound could have been obtained if the result (36) were not used. This, however, would have produced a more cumbersome bound. The $n = 0$ case may have been accommodated in the general n case since, as given in (9), $Q_0(x,t;w) = w(t)$ and since $w(t) > 0$ then $|Q_0(x,t;w)| = w(t)$.*

The following theorem gives bounds on $|T_n(a,x,b;f;w)|$ in terms of the Lebesgue norms of $f^{(n+1)}(t)$.

Theorem 7 *Let the general conditions of Theorem 5 hold and further, let $f^{(n)}(t)$ be absolutely continuous for $t \in [a,b]$, then*

$$|T_n(a,x,b;f;w)| \leq \begin{cases} [M_{n+1}(a,x;w) + M_{n+1}(x,b;w)] \left\|f^{(n+1)}\right\|_\infty, & f^{(n+1)} \in L_\infty[a,b]; \\ \|Q_n(x,\cdot;w)\|_q \left\|f^{(n+1)}\right\|_p, & f^{(n+1)} \in L_p[a,b], \\ & p>1,\ \frac{1}{p}+\frac{1}{q}=1; \\ [M_n(a,x;w) + M_n(x,b;w) \\ \quad + |M_n(a,x;w) - M_n(x,b;w)|]\dfrac{\left\|f^{(n+1)}\right\|_1}{2}, & f^{(n+1)} \in L_1[a,b], \end{cases} \tag{38}$$

where $M_n(a,x;w)$ and $M_n(x,b;w)$ are as given by (6) and (7), $Q_n(a,x,b;w)$ is defined in (9) and $T_n(a,x,b;w)$ by (12).

Further, $\|\cdot\|_p$ signify the usual Lebesgue norms where

$$\|h\|_\infty := ess \sup_{t\in[a,b]} |h(t)| \quad for \quad h \in L_\infty[a,b],$$

and

$$\|h\|_p := \left(\int_a^b |h(t)|^p\, dt\right)^{\frac{1}{p}} \quad for \quad h \in L_p[a,b],\ 1 \leq p < \infty.$$

Proof. From the identity (22) we have on using properties of the modulus and integral,

$$|T_n(a,x,b;w)| \leq \int_a^b \left|Q_{n+1}(x,t;w)\, f^{(n+1)}(t)\right| dt. \tag{39}$$

Now, for $f^{(n+1)} \in L_\infty[a,b]$

$$\int_a^b \left|Q_{n+1}(x,t;w)\, f^{(n+1)}(t)\right| dt \leq \left\|f^{(n+1)}\right\|_\infty \int_a^b |Q_{n+1}(x,t;w)|\, dt,$$

which upon using (33) produces the first inequality in (38).

For the second bound we use Hölder's integral inequality in (39) to give

$$\begin{aligned} &\int_a^b \left|Q_{n+1}(x,t;w)\, f^{(n+1)}(t)\right| dt \\ &\qquad \leq \left(\int_a^b |Q_{n+1}(x,t;w)|^q\, dt\right)^{\frac{1}{q}} \left(\int_a^b \left|f^{(n+1)}(t)\right|^p dt\right)^{\frac{1}{p}} \\ &\qquad = \|Q_n(x,\cdot;w)\|_q \left\|f^{(n+1)}\right\|_p, \end{aligned}$$

where $p > 1$, $\frac{1}{p} + \frac{1}{q} = 1$.

The final inequality in (38) is obtained from (39) to give

$$\int_a^b \left| Q_{n+1}(x,t;w) f^{(n+1)}(t) \right| dt \leq \sup_{t\in[a,b]} |Q_{n+1}(x,t;w)| \int_a^b \left| f^{(n+1)}(t) \right| dt,$$

where we may use (28) and a property of the $\max\{X,Y\}$ to obtain the stated result. ■

Remark 4 *If we take the weight function $w(t) \equiv 1$ then the results of Cerone et al. [7] involving the generalised trapezoidal rule and $n-$time differentiable functions is recaptured.*

A question that needs to be asked is can we choose the parameter x in such a way that the bound is minimized? The following lemma examines such an issue.

Lemma 4 *Let*

$$\begin{aligned} &2\phi_n(a,x,b;w) \\ &\quad = [M_n(a,x;w) + M_n(x,b;w) + |M_n(a,x;w) - M_n(x,b;w)|] \end{aligned} \tag{40}$$

and

$$\Psi_{n+1}(a,x,b;w) = M_{n+1}(a,x;w) + M_{n+1}(x,b;w). \tag{41}$$

Then

$$\phi_n^*(a,\tilde{x},b;w) = \min_{x\in[a,b]} \phi_n(a,x,b;w) = \frac{M_n(a,\tilde{x};w) + M_n(\tilde{x},b;w)}{2}, \tag{42}$$

where $\tilde{x}$ is the solution of $M_n(a,x;w) = M_n(x,b;w)$.

Further,

$$\begin{aligned} \Psi_{n+1}^*\left(a,\frac{a+b}{2},b;w\right) &= \min_{x\in[a,b]} \Psi_{n+1}(a,x,b;w) \\ &= M_{n+1}\left(a,\frac{a+b}{2};w\right) + M_{n+1}\left(\frac{a+b}{2},b;w\right), \end{aligned} \tag{43}$$

where $w(\cdot)$ is a positive weight function and $x \in [a,b]$, with $M_n(a,x;w)$ and $M_n(x,b;w)$ are as defined in (6) and (7).

Proof. The functions $M_n(a,x;w)$ and $M_n(x,b;w)$ are both positive with $M_n(a,x;w)$ increasing and $M_n(x,b;w)$ decreasing in $x \in [a,b]$. Thus, the minimum is attained when $|M_n(a,x;w) - M_n(x,b;w)| = 0$ giving the result as stated.

Now, $\Psi_{n+1}(a,x;w) \geq 0$ and

$$\Psi_{n+1}(a,a;w) = \Psi_{n+1}(b,b;w) = M_n(a,b;w).$$

Also, for $w(x) > 0$

$$\Psi'_{n+1}(a,x,b;w) = \left[(x-a)^n - (b-x)^n\right] \frac{w(x)}{n!}$$

and so $\Psi'_{n+1}(a,a,b;w) < 0$, $\Psi'_{n+1}(a,b,b;w) > 0$ bringing us to the conclusion that $\Psi_n(a,x,b;w)$ is convex in x. Since $w(x) > 0$, the minimum is attained when $x = \frac{a+b}{2}$, making $\Psi'_{n+1}\left(a, \frac{a+b}{2}, b; w\right) = 0$. ■

The following lemma obtains some coarser bounds which may prove to be more useful in practice. It involves obtaining bounds on

$$\Psi_{n+1}(a,x,b;w) = \|Q_{n+1}(x,\cdot;w)\|_1 = M_{n+1}(a,x;w) + M_{n+1}(x,b;w). \quad (44)$$

Lemma 5 *Let $w(t)$ be a weight function defined on $[a,b]$ and $x \in [a,b]$, then*

$$\begin{aligned} |\Psi_{n+1}(a,x,b;w)| &= \|Q_{n+1}(x,\cdot;w)\|_1 \\ &\leq \begin{cases} C(1)\|w\|_\infty, & w \in L_\infty[a,b]; \\ C^{\frac{1}{2}}(q)\|w\|_p, & w \in L_p[a,b], \\ & p > 1,\ \frac{1}{p} + \frac{1}{q} = 1; \\ \dfrac{\nu^{n+1}}{(n+1)!}\|w\|_1, & w \in L_1[a,b], \end{cases} \end{aligned} \quad (45)$$

where

$$C(q) = \frac{(x-a)^{q(n+1)+1} + (b-x)^{q(n+1)+1}}{q(n+1)+1}$$

and

$$\nu = \frac{b-a}{2} + \left|x - \frac{a+b}{2}\right|.$$

Proof. From the definitions (6) and (7) it may be noticed that $\Psi_{n+1}(a,x,b;w)$ from (44) may be expressed as

$$\begin{aligned} \Psi_{n+1}(a,x,b;w) &= \|Q_{n+1}(a,x,b;w)\|_1 \\ &= \frac{1}{(n+1)!}\int_a^b \kappa^{n+1}(a,x,b;u)\, w(u)\, du, \end{aligned} \quad (46)$$

where

$$\kappa(a,x,b;u) = \begin{cases} u-a, & u \in [a,x], \\ b-u, & u \in (x,b]. \end{cases} \quad (47)$$

Now,

$$(n+1)!\,|\Psi_{n+1}(a,x,b;w)| \leq \int_a^b \left|\kappa^{n+1}(a,x,b;u)\, w(u)\right| du$$

and so for $w \in L_p[a,b]$, $1 < p < \infty$ then

$$\begin{aligned} &\int_a^b \left|\kappa^{n+1}(a,x,b;u)\, w(u)\right| du \\ &\qquad \leq \left(\int_a^b \kappa^{q(n+1)}(a,x,b;u)\, du\right)^{\frac{1}{q}} \left(\int_a^b w^p(u)\, du\right)^{\frac{1}{p}}. \end{aligned} \quad (48)$$

Explicitly,

$$\left(\int_a^b \kappa^{q(n+1)}(a,x,b;u)\,du\right)^{\frac{1}{q}}$$
$$= \left[\int_a^x (u-a)^{q(n+1)}\,du + \int_x^b (b-u)^{q(n+1)}\,du\right]^{\frac{1}{q}}$$
$$= \left[\frac{(x-a)^{q(n+1)+1} + (b-x)^{q(n+1)+1}}{q(n+1)+1}\right]^{\frac{1}{q}}$$

which together with (48) gives the second inequality (45).

For $w \in L_\infty[a,b]$, then

$$\int_a^b \left|\kappa^{n+1}(a,x,b;u)\,w(u)\right| du \le C(1)\,\|w\|_\infty$$

the first inequality in (45).

Finally, for $w \in L_1[a,b]$, then

$$\int_a^b \left|\kappa^{n+1}(a,x,b;u)\,w(u)\right| du \le \left[\sup_{x\in[a,b]} \kappa(a,x,b;u)\right]^{n+1} \|w\|_1$$

where

$$\sup_{x\in[a,b]} \kappa(a,x,b;u) = \max\{x-a, b-x\} = \frac{b-a}{2} + \left|x - \frac{a+b}{2}\right| = \nu.$$

■

Karamata [12] proved the following theorem.

Theorem 8 *Let $g, w : [a,b] \to \mathbb{R}$ be integrable on $[a,b]$ and suppose $m \le g(t) \le M$ and $0 < c \le w(t) \le \lambda c$ for $t \in [a,b]$ and some constants m, M, c and λ. If G and $A(g,w)$ are defined as*

$$G := \frac{1}{b-a}\int_a^b g(t)\,dt \quad \text{and} \quad A(g,w) := \frac{\int_a^b g(t)\,w(t)\,dt}{\int_a^b w(t)\,dt} \tag{49}$$

then

$$\frac{\lambda m(M-G) + M(G-m)}{\lambda(M-G) + (G-m)} \le A(g,w) \le \frac{m(M-G) + \lambda M(G-m)}{(M-G) + \lambda(G-m)}. \tag{50}$$

Using the above theorem of Karamata, the third inequality in (45) may be improved.

If we associate $\kappa^{n+1}(a,x,b;w)$, as defined by (47), with $g(t)$ above, then

$$0 \le \kappa(a,x,b;u) \le \nu = \max\{x-a, b-x\} = \frac{b-a}{2} + \left|x - \frac{a+b}{2}\right|$$

and

$$G = \frac{C(1)}{b-a} = \frac{1}{b-a}\int_a^b \kappa^{n+1}(a,x,b;u)\,du.$$

Hence, from (46)

$$\begin{aligned}(n+1)!\Psi_{n+1}(a,x,b;w) &= \|Q_{n+1}(x,\cdot;w)\|_1 \\ &\le \frac{\lambda\nu^{n+1}C(1)\|w\|_1}{(b-a)\nu^{n+1} - C(1) + \lambda C(1)} \\ &\le \nu^{n+1}\|w\|_1.\end{aligned}$$

The last inequality follows from the fact that

$$C(1) = \int_a^b \kappa^{n+1}(a,x,b;u)\,du \le \nu^{n+1}(b-a).$$

4 Comparison of Ostrowski and Trapezoidal Results

The generalised weighted trapezoid kernel $Q_n(x,t;w)$ defined by (9) and (11) is a mapping $Q_n(\cdot,\cdot;w) : [a,b]^2 \to \mathbb{R}$, $a,b \in \mathbb{R}$, $a < b$ where

$$Q_n(x,t;w) := \begin{cases} \dfrac{(-1)^n}{(n-1)!}\displaystyle\int_t^x (u-t)^{n-1} w(u)\,du, & t \in [a,x], \quad n \in \mathbb{N} \\ \dfrac{(-1)^n}{(n-1)!}\displaystyle\int_x^t (t-u)^{n-1} w(u)\,du, & t \in (x,b], \quad n \in \mathbb{N} \\ w(t), & n = 0. \end{cases} \tag{51}$$

Here $w(t)$ is a weight function with properties as ascribed earlier in the paper. An identity relating the generalised weighted trapezoid functional $T_n(a,x,b;f;w)$ as defined by (12) is given by (13).

In our notation define the weighted Ostrowski functional $\Theta(a,x,b;f;w)$ by

$$\Theta_n(a,x,b;f;w) := \int_a^b w(t) f(t)\,dt - \sum_{k=0}^{n} E_k(a,x,b;w) f^{(k)}(x), \tag{52}$$

where

$$E_k(a,x,b;w) = \frac{1}{k!}\int_t^x (u-x)^k w(u)\,du. \tag{53}$$

Let the kernel $K_n(x,t;w)$ be such that $K_n(\cdot,\cdot,w) : [a,b]^2 \to \mathbb{R}$, $a,b \in \mathbb{R}$,

$a < b$ and $w(\cdot)$ a given weight function, then

$$K_n(x,t;w) := \begin{cases} \dfrac{1}{(n-1)!}\displaystyle\int_a^t (t-u)^{n-1} w(u)\,du, & t \in [a,x), \quad n \in \mathbb{N} \\ 0 & t = x \\ \dfrac{(-1)^n}{(n-1)!}\displaystyle\int_t^b (u-t)^{n-1} w(u)\,du, & t \in (x,b], \quad n \in \mathbb{N} \\ w(t), & x,t \in [a,b]. \end{cases} \tag{54}$$

Matić et al. [13] show that for $f^{(n)}(\cdot)$ continuous and of bounded variation on $[a,b]$ then for $x \in [a,b]$

$$\Theta_n(a,x,b;f;w) = (-1)^{n+1}\int_a^b \kappa^{n+1}(x,t;w)\,df^{(n)}(t). \tag{55}$$

Further, for $f^{(n)}(\cdot)$ absolutely continuous on $[a,b]$ then $df^{(n)}(t) = f^{(n+1)}(t)\,dt$, giving the identity

$$\Theta_n(a,x,b;f;w) = (-1)^{n+1}\int_a^b \kappa^{n+1}(x,t;w)\,f^{(n+1)}(t)\,dt \tag{56}$$

from (55).

The bounds for $|T(a,x,b;f;w)|$ and $|\Theta(a,x,b;f;w)|$ depend on the behaviour of $|Q_{n+1}(x,t;w)|$ and $|K_{n+1}(x,t;w)|$ respectively.

The following lemma gives sufficient conditions for the bounds on $|T(a,x,b;f;w)|$ and $|\Theta(a,x,b;f;w)|$ to be equal.

Lemma 6 *For $w : [a,b] \to (0,\infty)$, $\int_a^b w(t)\,dt < \infty$ and $w(t)$ symmetric about the respective midpoints for $t \in [a,x)$ and $t \in (x,b]$ then,*

$$|K_n(x,t;w)| = \begin{cases} |Q_n(x,a+x-t;w)|, & t \in [a,x), \\ 0 & t = x \\ |Q_n(x,x+b-t;w)|, & t \in (x,b]. \end{cases} \tag{57}$$

An interchange of Q_n and K_n in (57) is valid under the same conditions.

Proof. Consider for $t \in [a,x)$

$$\begin{aligned} (n-1)!\,|Q_n(x,a+x-t;w)| &= \int_{a+x-t}^{x} [u-(a+x-t)]^{n-1} w(u)\,du \\ &= \int_a^t (t-v)^{n+1} w(a+x-v)\,dv \\ &= (n-1)!\,|K_n(x,t;w)|, \end{aligned}$$

provided $w(a+x-v) = w(v)$, that is, $w\left(\frac{a+x}{2} - z\right) = w\left(\frac{a+x}{2} + z\right)$ for $z \in [a, x)$. A similar argument gives the result for $t \in (x, b]$. We note that if t varies from c to d, then $T = c + d - t$ varies from d to c. Thus the interchange of K_n and Q_n in (57) is valid and the equivalent expression to (57) holds. ■

Remark 5 *A consequence of Lemma 6 is that if $w(t) \equiv \alpha$, a constant, then the Lebesgue norms giving the bounds for $|T(a, x, b; f; w)|$ and $|\Theta(a, x, b; f; w)|$ are equal. In particular, the unweighted case $w(t) = \alpha = 1$ produce the same bounds as shown in Cerone [4]. Weights such as*

$$w_n(t) = \begin{cases} \left(t - \frac{a+x}{2}\right)^{2n}, & t \in [a, x) \\ 0, & t = x \\ \left(t - \frac{x+b}{2}\right)^{2n}, & t \in (x, b], \quad n \in \mathbb{N} \end{cases}$$

and

$$W(t) = \begin{cases} \left|t - \frac{a+x}{2}\right|, & t \in [a, x) \\ 0, & t = x \\ \left|t - \frac{x+b}{2}\right|, & t \in (x, b], \end{cases}$$

would produce the same bounds for the trapezoidal and Ostrowski functions defined by (13) and (52).

Acknowledgement 1 *The work for the paper was undertaken while the first named author was on sabbatical at La Trobe University, Bendigo.*

References

[1] R.P. AGARWAL and S.S. DRAGOMIR, An application of Hayashi's inequality for differentiable functions, *Computers Math. Appl.,* **32**(6) (1996), 95-99.

[2] G.A. ANASTASSIOU, Ostrowski type inequalities, *Proc. Amer. Math. Soc.*, **123**(12) (1995), 3755-3781.

[3] G.A. ANASTASSIOU, Multivariate Ostrowski type inequalities, *Acta. Math.Hungar.*, **76** (1997), 267-278.

[4] P. CERONE, On relationships between Ostrowski, Trapezoidal and Chebychev identities and inequalities, submitted to *Soochow J. of Math.,* [ONLINE] `http://rgmia.vu.edu.au/v4n2.html`

[5] P. CERONE and S. S. DRAGOMIR, Lobatto type quadrature rules for functions with bounded derivative, *Math. Ineq. & Appl.,* **3**(2) (2000), 197-209.

[6] P. CERONE and S. S. DRAGOMIR, On a weighted generalisation of Iyengar type inequalities involving bounded first derivative, *Math. Ineq. & Appl.,* **3**(1) (2000), 35-44.

[7] P. CERONE, S. S. DRAGOMIR, J. ROUMELIOTIS, and J. SUNDE, A new generalisation of the trapezoid formula for $n-$time differentiable mappings and applications, *Demonstratio Math.,* **33**(4) (2000), 719-736.

[8] S.S. DRAGOMIR, P. CERONE, J. ROUMELIOTIS and S. WANG, A weighted version of Ostrowski inequality for mappings of Hölder type and applications in numerical analysis, *Bull. Math. Soc. Sc. Math. Roumanie,* **42**(90)(4) (1992), 301-304.

[9] S.S. DRAGOMIR and T.M. RASSIAS (Ed.), Ostrowski Type Inequalities and Applications in Numerical Integration, *Accepted for publication by Kluwer Academic Publishers,* Preprint available at: `http://rgmia.vu.edu.au/monographs/Ostrowski.html`

[10] S.S. DRAGOMIR and S. WANG, Applications of Iyenar's type inequalities to the estimation of error bounds for the trapezoidal quadrature rule, *Tamkang J. of Math.*, **29**(1) (1998), 55-58.

[11] K.S.K. IYENGAR, Note on an inequality, *Math. Student,* **6** (1938), 75-76.

[12] J. KARAMATA, O prvom stavu srednjih vrednosti odredjenih integrala, Glas srpske kraljevske akademije CLIV, Beograd (1933), 119-144.

[13] M. MATIĆ, J.E. PEČARIĆ and N. UJEVIĆ, Generalizations of weighted version of Ostrowski's inequality and some related results, *J. of Ineq. & App.,* **5** (2000), 639-666.

[14] G.V. MILOVANOVIĆ, On some integral inequalities, *Univ. Beograd Publ. Elektrotehn Fak. Ser. Mat. Fiz.,* No. 498-541 (1975), 119-124.

[15] G.V. MILOVANOVIĆ, O nekim funkcionalnim nejednakostima, *Univ. Beograd Publ. Elektrotehn Fak. Ser. Mat. Fiz.,* No. 599 (1977), 1-59.

[16] G.V. MILOVANOVIĆ and J.E. PEČARIĆ, On generalisation of the inequality of A. Ostrowski and some related applications, *Univ. Beograd Publ. Elektrotehn Fak. Ser. Mat. Fiz.,* No. 544-576.

[17] D.S. MITRINOVIĆ, J.E. PEČARIĆ and A.M. FINK, *Classical and New Inequalities in Analysis,* Kluwer Academic Publishers, Dordrecht, 1993.

[18] D.S. MITRINOVIĆ, J.E. PEČARIĆ and A.M. FINK, *Inequalities for Functions and Their Integrals and Derivatives,* Kluwer Academic Publishers, Dordrecht, 1994.

[19] A. OSTROWSKI, Uber die Absolutabweichung einer differentienbaren Funktionen von ihren Integralmittelwert, *Comment. Math. Hel,* **10** (1938), 226-227.

[20] C.E.M. PEARCE and J. PEČARIĆ, OnAnastassiou's generalisations of the Ostrowski inequality and related results, *J. Comput. Anal. and Applic.,* **3** (1) (2000), 25-34.

[21] FENG QI, Inequalities for a weighted integral, *RGMIA Research Report Collection,* **2**(7) (1999), Article 2. `http://rgmia.vu.edu.au/v2n7.html`

[22] FENG QI, Inequalities for a weighted multiple integral, *J. Math. Anal. & Appl.,* (in press). *RGMIA Research Report Collection,* **2**(7) (1999), Article 4.

[23] J. ROUMELIOTIS, P. CERONE and S.S. DRAGOMIR, An Ostrowski type inequality for weighted mappings with bounded second derivatives, *J. KSIAM,* **3**(2) (1999), 107–119.

[24] P.M. VASIĆ and G.V. MILOVANOVIĆ, On an inequality of Iyengar, *Univ. Beograd. Publ. Elektrotehn, Fak. Ser. Mat. Fiz.,* No. 544-576 (1976), 18-24. `http://rgmia.vu.edu.au/v2n7.html`

JOURNAL OF CONCRETE AND APPLICABLE MATHEMATICS,VOL.3,NO.4,457-468,
2005,COPYRIGHT 2005 EUDOXUS PRESS,LLC

Parallel schemes for two-dimensional parabolic equation with a boundary integral condition

Mehdi Dehghan
Department of Applied Mathematics,
Faculty of Mathematics and Computer Science,
Amirkabir University of Technology, No. 424 Hafez Avenue, Tehran, Iran

Abstract- Several numerical schemes are developed for obtaining approximate solutions to an initial boundary-value problem for the second-order parabolic partial differential equations (PDEs) with an integral condition replacing one boundary condition. The space derivatives in the PDE are approximated by finite difference replacements. The solution of the resulting system of first-order ordinary differential equations (ODEs) satisfies a recurrence relation which involves a matrix exponential function. The accuracy in time is controlled by choosing several subdiagonal Pade approximants to replace this matrix exponential term. Numerical techniques are developed to compute the required solution using a splitting method, leading to algorithms for sequential and parallel implementation. The algorithms are tested on a model problem from the literature. The central processor unit (CPU) times needed are also considered and are compared .

KEYWORDS: Two-Dimensional Diffusion-Numerical Integration- Sequential and Parallel Algorithms- Stability-Pade Approximant-Boundary Integral Condition- Central Processor Time- Partial Differential Equations.

1 Introduction

Two-dimensional parabolic equation with an integral condition replacing one boundary condition arises in many important applications in heat transfer [1, 2, 3, 4, 5, 6, 7], control theory [8], medical science [9], and thermoelasticity [10, 11, 12]. So, recently much attention has been given in the literature to the developement, analysis and implementation of accurate methods for the numerical solution of time-dependent partial differential equations with a boundary integral condition.

The purpose of this article is to present very efficient parallel methods for solving the following two-dimensional time-dependent diffusion equation

$$\frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2}, \tag{1}$$

with initial condition

$$u(x, y, 0) = f(x, y), \quad 0 \leq x, y \leq 1, \tag{2}$$

and boundary conditions

$$u(0,y,t) = g_0(y,t), \quad 0 \le t \le T, \quad 0 \le y \le 1, \tag{3}$$

$$u(1,y,t) = g_1(y,t), \quad 0 \le t \le T, \quad 0 \le y \le 1, \tag{4}$$

$$u(x,0,t) = h_0(x)\mu(t), \quad 0 \le t \le T, \quad 0 \le x \le 1, \tag{5}$$

$$u(x,1,t) = h_1(x,t), \quad 0 \le t \le T, \quad 0 \le x \le 1, \tag{6}$$

and the integral condition

$$\int_0^1 \int_0^{s(x)} u(x,y,t)dxdy = m(t), \quad 0 \le x,y \le 1, \tag{7}$$

where f, g_0, g_1, h_0, h_1, s and m are known functions, while the functions u and μ are to be determined.

The boundary condition (5) is variable separable, with spatial dependence given by $h_0(x)$ and time dependence given by $\mu(t)$.

The existence and uniqueness of the solution of this non-classic problem has been studied in [3].

An overview of this paper is as follows:

Numerical schemes for the solution of (1)-(7) are described in Section 2. Developement of numerical methods which will be based on the Pade approximants is presented in Section 3. A discussion on the sequential and parallel algorithms is given in Section 4. The method of incorporating (7) with μ unknown is described in Section 5, and numerical results for a test problem produced by the method developed, are given in Section 6. Section 7 concludes this paper with a brief summary.

2 The Numerical Solution with Dirichlet Boundary Conditions

We divide the domain $[0,1]^2 \times [0,T]$ into an $M^2 \times N$ mesh with spatial step size $h = 1/M$ in both x and y directions and the time step size $k = T/N$ respectively. Grid points (x_i, y_j, t_n) are given by

$$x_i = ih, \quad i = 0,1,2,\ldots,M, \tag{8}$$

$$y_j = jh, \quad j = 0,1,2,\ldots,M, \tag{9}$$

$$t_n = nl, \quad n = 0,1,2,\ldots,N, \tag{10}$$

in which M is an even integer. We use $u_{i,j}^n$ and μ^n to denote the approximations of $u(ih, jh, nl)$ and $\mu(nl)$, respectively.

The solution vector U^n will be ordered in the following form:

$$U^n = (u_{1,1}^n, u_{2,1}^n, \ldots, u_{M-1,1}^n, u_{1,2}^n, u_{2,2}^n, \ldots, u_{M-1,2}^n, \ldots, u_{1,M-1}^n, \ldots, u_{M-1,M-1}^n)^T. \tag{11}$$

The numerical methods suggested here are based on two ideas: Firstly, the method of lines semi-discretization approach will be used to transform the model partial differential

equation into a system of first-order linear ordinary differential equations, the solution of which satisfies a certain recurrence relation involving matrix exponential terms. The developement of numerical methods will be based on Pade approximations to such exponentials. A suitable rational approximant will be used to approximate such exponentials leading to an algorithm which may be parallelized through a partial-fraction splitting technique. This technique is used to approximate the solution of the two-dimensional diffusion equation, at interior grid points. Secondly, a highly accurate numerical integration scheme [13] is used to approximate the unknown function $\mu(t)$, using the integral condition (7).

The space derivative in (1) will be replaced by their second-order central difference approximation given by

$$\frac{\partial^2 u}{\partial x^2} = h^{-2}(u(x-h,y,t) - 2u(x,y,t) + u(x+h,y,t)) + O(h^2), \tag{12}$$

$$\frac{\partial^2 u}{\partial y^2} = h^{-2}(u(x,y-h,t) - 2u(x,y,t) + u(x,y+h,t)) + O(h^2), \tag{13}$$

as $h \to 0$.

Applying (1) to all the $(M-1)^2$ interior mesh points of the square at $[0,1] \times [0,1]$, at time level $t_n = nl$, with the space derivatives replaced by (12) and (13), leads to a system of $(M-1)^2$ first-order linear ordinary differential equations of the form [16]

$$\frac{dU(t)}{dt} = AU(t) + \psi(t), \quad t > 0, \tag{14}$$

with initial condition

$$U(0) = f, \tag{15}$$

in which the matrix A is of order $(M-1)^2$ and will be split into the constituent matrices B,C which commute such that $A = B + C$.

The vector $\psi(t)$, of order $(M-1)^2$, arises from the use of the time dependent boundary conditions (3)-(6) in (12) and (13).

The matrix B arises from the use of (12) in (1), it is block diagonal with tridiagonal blocks and has the form:

$$B = h^{-2}\begin{bmatrix} Q & 0 & \dots & & & 0 \\ 0 & Q & 0 & \dots & & \vdots \\ \vdots & 0 & Q & 0 & \dots & \\ & \vdots & & & & \\ & & & & Q & 0 \\ 0 & \dots & & & 0 & Q \end{bmatrix}, \tag{16}$$

where Q is the tridiagonal matrix of order $M-1$ given by

$$Q = \begin{bmatrix} -2 & 1 & 0 & \dots & & & 0 \\ 1 & -2 & 1 & 0 & \dots & & \vdots \\ \vdots & 1 & -2 & 1 & \dots & & \\ & \vdots & & & & & \\ & & & & & -2 & 1 \\ 0 & \dots & & & & 1 & -2 \end{bmatrix}. \tag{17}$$

The matrix C arises from the use of (13) in (1), it is block tridiagonal with diagonal blocks and has the form:

$$C = h^{-2} \begin{bmatrix} -2I & I & 0 & \dots & & & 0 \\ I & -2I & I & 0 & \dots & & \vdots \\ \vdots & I & -2I & I & \dots & & \\ & \vdots & & & & & \\ & & & & & -2I & I \\ 0 & \dots & & & 0 & I & -2I \end{bmatrix}, \tag{18}$$

where I is the identity matrix of order $M-1$.

The $(M-1)^2$ eigenvalues of the matrix A are real and negative and are given by

$$\lambda_{i,j} = -4h^{-2}[sin^2\frac{i\pi}{2M} + sin^2\frac{j\pi}{2M}], \quad i,j = 1,2,\dots,M-1. \tag{19}$$

Solving the system of ODEs (14) subject to the initial condition (15) gives

$$U(t) = exp(tA)f + \int_0^t exp[(t-s)A]\psi(s)ds, \quad t \geq 0, \tag{20}$$

which satisfies the recurrence relation [17]

$$U(t+l) = exp(lA)U(t) + \int_t^{t+l} exp[(t+l-s)A]\psi(s)ds, \quad t = 0, l, 2l, \dots \tag{21}$$

in which l is a constant time step in the discretization of the time variable $t \geq 0$, at the points $t_n = nl(n = 0, 1, 2, \dots)$.

3 Solution at the First Time Step and the Pade Approximant

Following [17] and using the trapezoidal rule for evaluating the quadrature in (21), gives the following second-order formula:

$$U(t+l) = exp(lA)U(t) + \frac{l}{2}[\psi(t+l) + exp(lA)\psi(t)], \quad t = 0, l, 2l, 3l, \dots \tag{22}$$

Replacing the matrix A by $B + C$ and using the commutativity of B and C, (22) may be written as

$$U(t+l) = exp(lB)exp(lC)U(t) + \frac{l}{2}[\psi(t+l) + exp(lB)exp(lC)\psi(t)], \quad t = 0, l, 2l, 3l, 4l, \ldots. \tag{23}$$

The developement of the numerical methods will be based on making appropriate approximations in this recurrence relation. Higher-order Pade approximants are popularly employed for such exponentials. Methods based on the use of these approximants are of high accuracy in time and, in the case of the subdiagonal Pades, have good stability properties.

The pade approximant to the exponential function e^θ, where θ is a real scalar, has the form

$$e^\theta \approx R_{b,d}(\theta) = \frac{P_d(\theta)}{Q_b(\theta)}, \tag{24}$$

where $P_d(\theta)$ and $Q_b(\theta)$ are polynomials of degrees d and b, respectively with real coefficients, in each of which the constant term is unity. The polynomials $p_d(\theta)$ and $Q_b(\theta)$, for the approximants to be used in Section 4 are given in Table 1.

$(b,d)Pade$	$P_d(\theta)$	$Q_b(\theta)$
$(2,0)$	1	$1-\theta+\frac{\theta^2}{2}$
$(2,1)$	$1+\frac{\theta}{3}$	$1-\frac{2\theta}{3}+\frac{\theta^2}{6}$
$(3,0)$	1	$1-\theta+\frac{\theta^2}{2}-\frac{\theta^3}{6}$
$(2,2)$	$1+\frac{\theta}{2}+\frac{\theta}{12}$	$1-\frac{\theta}{2}+\frac{\theta^2}{12}$
$(2,3)$	$1+\frac{3\theta}{5}+\frac{3\theta^2}{20}+\frac{\theta^3}{60}$	$1-\frac{2\theta}{5}+\frac{\theta^2}{20}$
$(2,4)$	$1+\frac{2\theta}{3}+\frac{\theta^2}{5}+\frac{\theta^3}{30}+\frac{\theta^4}{360}$	$1-\frac{\theta}{3}+\frac{\theta^2}{30}$

Table 1: Pade approximant to e^θ

Note that the [d,b] denote the Pade approximation to e^θ. High-order stable subdiagonal ($b > d$) Pade methods can be generated which are particularly efficient on machines utilizing several concurrent processors.

It will be assumed that $Q_b(\theta)$ has b_1 real zero and $2b_2$ complex zeros occuring in complex conjugate pairs so that $b = b_1 + 2b_2$. So $R_{b,d}(\theta)$ can be expanded in partial fraction form, resulting in approximating the matrix exponential function in (23) by

$$exp(A) \simeq \sum_{i=1}^{b_1} w_i(A - c_i I)^{-1} + 2\sum_{i=b_1+1}^{b_1+b_2} Re[w_i(A - c_i I)]^{-1}. \tag{25}$$

The above expansion is particularly useful in parallel computing environments because it can be used to apportion the work of solving the corresponding linear algebraic systems to processors operating concurrently. Consequently, the solution of (14) can be implemented on a machine with $b_1 + b_2$ processors through parallel implicit Euler-like schemes, as described in Section 4. We can also construct parallel algorithms based on diagonal $[b, b]$ Pade approximations.

The numerical method to be developed in this report is based on the use of the subdiagonal [1/2] Pade approximant given by

$$exp(lA) \approx (I - \frac{2}{3}lA + \frac{1}{6}l^2A^2)^{-1}(I + \frac{1}{3}lA), \tag{26}$$

for the matrix exponentials in the recurrence relation (23) to give the $O(h^2 + l^3)$ unconditionally stable scheme [18],

$$\begin{aligned} u(t+l) &= (I - \frac{2}{3}lB + \frac{1}{6}l^2B^2)^{-1}(I + \frac{1}{3}lB) \\ &\quad (I - \frac{2}{3}lC + \frac{1}{6}l^2C^2)^{-1}(I + \frac{1}{3}lC)U(t) \\ &+ \frac{l}{2}[(I - \frac{2}{3}lB + \frac{1}{6}l^2B^2)^{-1}(I + \frac{1}{3}lB) \\ &\quad (I - \frac{2}{3}lC + \frac{1}{6}l^2C^2)^{-1}(I + \frac{1}{3}lC)\psi(t) + \psi(t+l)]. \end{aligned} \tag{27}$$

4 Sequential and Parallel Algorithms

The implementation of the sequential algorithms resulting from pre-multiplying the matrix inverses in (27) involves using higher powers of the block matrices B and C, thus requiring considerable amount of CPU time and computer storage as reported in [14]. We can overcome this difficulty using the parallel algorithms. The expansion (25) has great importance in the parallel computing environment in that it can be used to solve a corresponding linear algebraic system on processors operating concurrently. Using this expansion, parallel algorithms may be implemented on machines with $b_1 + 2b_2$ processors (only).

The partial-fraction of $(I - \frac{2}{3}lB + \frac{1}{6}l^2B^2)^{-1}(I + \frac{1}{3}lB)$ leads to :

$$(I - \frac{2}{3}lB + \frac{1}{6}l^2B^2)^{-1}(I + \frac{1}{3}lB) \simeq [k_1(I - z_1lB)^{-1} + k_2(I - z_2lB)^{-1}], \tag{28}$$

So we can express (27) in its partial-fraction splitting form [18]

$$\begin{aligned} U(t+l) &= [k_1(I - z_1lB)^{-1} + k_2(I - z_2lB)^{-1}][k_1(I - z_1lC)^{-1} \\ &+ k_2(I - z_2lC)^{-1}]U(t) + \frac{l}{2}\psi(t+l) \\ &+ \frac{l}{2}[k_1(I - z_1lB)^{-1} + k_2(I - z_2lB)^{-1}] \\ &\quad [k_1(I - z_1lC)^{-1} + k_2(I - z_2lC)^{-1}]\psi(t), \end{aligned} \tag{29}$$

where z_1 and z_2 are the poles of the $(2,1)$ Pade approximant and are given by:

$$z_1 = \bar{z}_2 = 0.33333333272613 + 0.23570226861019i, \tag{30}$$

and k_1 and k_2 are the constants resulting from the partial fraction decomposition and are given by:

$$k_1 = \bar{k_2} = \frac{1}{2} - \sqrt{2}i, \quad i = \sqrt{-1}. \tag{31}$$

Note that the closed form expressions for z_1 and z_2 are not available, so we have used their approximate values (30).

$U(t+l)$ in (29), the solution vector at time $t=(n+1)l$, may now be obtained via the parallel algorithms using two processors [18]:

Processor 1 :

$$\begin{aligned}(I-z_1lC)w_1 &= k_1U(t)\\ p_1 &= 2Re(w_1)\\ (I-z_1lB)w_2 &= k_1p_1\\ p_2 &= 2Re(w_2),\end{aligned} \tag{32}$$

Processor 2 :

$$\begin{aligned}(I-z_1C)w_3 &= k_1\psi(t)\\ p_3 &= 2Re(w_3)\\ (I-z_1lB)w_4 &= k_1p_3\\ p_4 &= 2Re(w_4),\end{aligned} \tag{33}$$

$$U(t+l) = p_2 + \frac{l}{2}[p_4+\psi(t+l)]. \tag{34}$$

The intermediate vectors $p_i, w_i, i=1,2,3,4$, need not to be stored once $U(t+l)$ is computed at each time step. The matrices $(I-z_1lB)$ and $(I-z_1lC)$ are decomposed into lower and upper triangular (LU) forms only once. These LU products are then "fed" to the two processors in order to compute the intermediate vectors $z_i, i=1,2,3,4$, using forward and backward substitutions.

The principal part of the local truncation error arising from the use of the (b,d) Pade approximant is

$$C_{q+1}l^q(\frac{\partial^{q+1}u}{\partial t^{q+1}}) - \frac{1}{12}h^2[(\frac{\partial^4u}{\partial x^4}) + (\frac{\partial^4u}{\partial y^4})], \tag{35}$$

$q=b+d$, where the error constants C_{q+1} for a selection of second, third and fourth-order methods (in time) are given in Table 2.

$order(q=b+d)$	$Pade$	$Error\quad Constants\quad (C_{q+1})$
2	$(1,1)$	$\frac{-1}{12}$
2	$(2,0)$	$\frac{1}{6}$
3	$(2,1)$	$\frac{1}{72}$
3	$(3,0)$	$\frac{-1}{24}$
4	$(3,1)$	$\frac{1}{-480}$
4	$(4,1)$	$\frac{1}{120}$

Table 2: The error constants C_q

5 The Computation of $\mu(t)$

In this section the numerical integration procedure to be used to approximate the unknown function $\mu(t)$ using the boundary integral condition (7) will be described.

The presence of an integral term in a boundary condition can greatly complicate the application of standard numerical techniques. The accuracy of the quadrature must be compatible with that of the discretization of the differential equation.

In the following the double integral in (7) is approximated using the fourth-order Simpson's composite "one-third "rule.

Consider the integral

$$H(x,t)=\int_0^{s(x)} u(x,y,t)dy. \tag{36}$$

Then application of Simpson's composite "one-third " rule gives

$$\int_0^1 H(x,t)dx \simeq \frac{h}{3}\Big(H_0+4\sum_{i=1}^{M/2} H_{2i-1}+2\sum_{i=1}^{(M/2)-1} H_{2i}+H_M\Big), \tag{37}$$

in which

$$H_i=\int_0^{2l_ih} u(x_i,y,t)dy+\int_{2l_ih}^{s(ih)} u(x_i,y,t)dy, \tag{38}$$

where

$$l_i=[s(ih)/2h], \tag{39}$$

and $[\,\cdot\,]$ represents the integer part of the argument.

Substituting in the second integral of (38)

$$z_i=y/h-2l_i \tag{40}$$

yields

$$\int_{2l_ih}^{s(ih)} u(x_i,y,t)dy=h\int_0^{\delta_i} u(x_i,z_i,t)dz_i, \tag{41}$$

where

$$\delta_i=(s(ih)/h)-2l_i. \tag{42}$$

Replacement of u in the integral with a quadratic interpolating polynomial(the Newton's forward-difference formula) through the grid values concerned, gives

$$\int_0^{\delta_i} u(x_i,z_i,t)dz_i=\int_0^{\delta_i}[u_{i,2l_i}+z_i\triangle u_{i,2l_i}+\frac{1}{2}z_i(z_i-1)\triangle^2 u_{i,2l_i}]dz_i+O(h^4), \tag{43}$$

where

$$\triangle u_{i,2l_i}=u_{i,2l_i+1}-u_{i,2l_i}, \tag{44}$$

and

$$\triangle^2 u_{i,2l_i}=u_{i,2l_i+2}-2u_{i,2l_i+1}+u_{i,2l_i}. \tag{45}$$

Integrating (43) and collecting the like terms means that

$$\begin{aligned} H_i \;\simeq\; & \frac{h}{3}[u_{i,0}+4\sum_{j=1}^{l_i} u_{i,2j-1}+2\sum_{j=1}^{l_i-1} u_{i,2j}+u_{i,2l_i}+3\delta_i(1-3\delta_i/4 \\ + \;& {\delta_i}^2/6)u_{i,2l_i}+3{\delta_i}^2(1-\delta_i/3)u_{i,2l_i+1}+({\delta_i}^2/4)(2\delta_i-3)u_{i,2l_i+2}]. \end{aligned} \tag{46}$$

Putting

$$V(t) = \int_0^1 H(x,t)dx, \tag{47}$$

and using the approximation

$$v^n \simeq \frac{h}{3}\Big(H_0^n + 4\sum_{i=1}^{M/2} H_{2i-1}^n + 2\sum_{i=1}^{(M/2)-1} H_{2i}^n + H_M^n\Big), \tag{48}$$

then gives

$$v^n \simeq \frac{h^2}{9}\Big(\sum_{i=0}^{(M/2)-1} u_{0,0}^n + 4\sum_{i=1}^{M} u_{2i-1,0}^n + 2\sum_{i=1}^{(M/2)-1} u_{2i,0}^n + u_{M,0}^n\Big) + R^n. \tag{49}$$

Note that

$$v^n \simeq \frac{h}{2}\mu^n \int_0^1 s(x)dx + R^n, \tag{50}$$

where R^n is the summation in v^n excluding the values at the boundary $y = 0$.

We then approximate μ^{n+1} by means of

$$\mu^{n+1} = \frac{m^{n+1} - R^n}{\frac{h}{3}\int_0^1 s(x)dx}, \qquad \int_0^1 s(x)dx \neq 0, \tag{51}$$

from which the boundary values along $y = 0$ may be computed at time t_{n+1} using the boundary condition (5).

6 Numerical Results

The numerical method described in the previous section was applied to the following problem:

Consider (1)–(7) with

$$f(x,y) = exp(x+y), \tag{52}$$

$$g_0(y,t) = exp(y+2t), \tag{53}$$

$$g_1(y,t) = exp(1+y+2t), \tag{54}$$

$$h_0(x) = exp(x), \tag{55}$$

$$h_1(x) = exp(1+x+2t), \tag{56}$$

$$\mu(t) = exp(2t), \tag{57}$$

$$m(t) = (4exp(exp(1)/4) - 4exp(1/4) - exp(1) + 1)exp(2t), \tag{58}$$

$$s(x) = exp(x)/4, \tag{59}$$

for which the exact solution is

$$u(x,y,t) = exp(x+y+2t). \tag{60}$$

			New Scheme	BTCS Scheme
x	y	Exact u	Error	Error
0.1	0.1	9.025013	-5.0×10^{-4}	-2.0×10^{-3}
0.2	0.2	11.023176	-6.0×10^{-4}	-3.0×10^{-3}
0.3	0.3	13.463738	-6.0×10^{-4}	-2.0×10^{-3}
0.4	0.4	16.444647	-5.0×10^{-4}	-3.0×10^{-3}
0.5	0.5	20.905243	-6.0×10^{-4}	-4.0×10^{-3}
0.6	0.6	24.532530	-7.0×10^{-4}	-3.0×10^{-3}
0.7	0.7	29.964100	-8.0×10^{-4}	-3.0×10^{-3}
0.8	0.8	36.598234	-6.0×10^{-4}	-2.0×10^{-3}
0.9	0.9	40.447304	-6.0×10^{-4}	-4.0×10^{-3}

Table 3: Results for u with $T = 1.0$, $h = 0.02$, $l = 0.0001$

The results for $u_{i,j}^N$ with $h = 0.02$, $l = \frac{1}{10000}$ at $T = 1.0$, using the method discussed in Section 4 and the (5,1) fully implicit finite difference BTCS formula of [3] are shown in Table 3.

The CPU time for the new technique was 53.1 sec while the CPU time for the fully implicit finite difference method of [3] was 1553 sec.

The results obtained for μ with $h = 0.02$, $l = \frac{1}{10000}$, using the new method developed in this paper and the (5,1) fully implicit finite difference BTCS technique to solve the example non-classic boundary value problem are shown in Table 4.

		New Scheme	BTCS Scheme
t	Exact μ	Error	Error
0.1	1.221403	7.0×10^{-4}	8.0×10^{-3}
0.2	1.491825	6.0×10^{-4}	7.0×10^{-3}
0.3	1.822119	6.0×10^{-4}	7.0×10^{-3}
0.4	2.225541	6.0×10^{-4}	6.0×10^{-3}
0.5	2.718282	7.0×10^{-4}	5.0×10^{-3}
0.6	3.320117	7.0×10^{-4}	6.0×10^{-3}
0.7	4.055200	8.0×10^{-4}	6.0×10^{-3}
0.8	4.953032	7.0×10^{-4}	5.0×10^{-3}
0.9	6.049647	6.0×10^{-4}	4.0×10^{-3}
1.0	7.389056	5.0×10^{-4}	5.0×10^{-3}

Table 4: Results for μ with $h = 0.02$, $l = 0.0001$

Note that the results showed that the new technique is about thirty times faster than the (5,1) fully implicit finite difference BTCS scheme.

7 Summary

In this paper a parallel algorithm was developed and applied to the two-dimensional

parabolic equation with a boundary integral condition replacing one boundary condition. The algorithm, which may be implemented on a parallel architecture using two processors, requires only the applications of tridiagonal solvers at every time step. This technique worked very well for two dimensional diffusion with an integral condition. For the model problems considered, the parallel algorithm (34) was found to be about thirty times faster than the fully implicit finite difference BTCS scheme of [14]. A comparison with the standard explicit finite difference scheme for the model problem clearly demonstrates that the new technique is computationally superior. The numerical results obtained by the new scheme discussed in this article on a test problem from the literature, give acceptable results and suggest convergence to the exact solution when h goes to zero. The new method discussed in this article is readily adaptable to similar three dimensional problems.

References

[1] J. R. CANNON and J. van der HOEK: Diffusion subject to specification of mass. *J. Math. Anal. Appl.* **115**, 517-529, (1986).

[2] J. R. CANNON: The solution of the heat equation subject to the specification of energy. *Q. Appl. Math.* **21**, 155-160 (1963).

[3] J. R. CANNON, Y. LIN and A. L. MATHESON: The solution of the diffusion Equation in two-space variables subject to the specification of mass. *Appl. Anal. J.* **50**, 1-19, (1993).

[4] J. R. CANNON and Y. LIN and S. WANG: Determination of a control function in a paraboilc partial differential equations, Research report, 89-10, Department of Mathematics and Statistics, Mcgill University.

[5] J. R. CANNON and H. M. YIN : On a class of non-classical parabolic problems, *J. Differential Equations*, **79**, 266-288, (1989).

[6] S. WANG and Y. LIN: A numerical method for the diffusion equation with nonlocal boundary specifications. *Int. J. Engng. Sci.* **28**, 543-546, (1990).

[7] S. WANG : The numerical method for the conduction subject to moving boundary energy specification, *Numerical Heat Transfer*, Vol. **130**, 35-38, (1990).

[8] S. WANG and Y. LIN: A finite difference solution to an inverse problem determining a control function in a parabolic partial differential equations. *Inverse Problems* **5**, 631-640, (1989).

[9] V. CAPSSO and K. KUNISCH: A reaction-diffusion system arising in modeling man-environment diseases. *Quart. Appl. Math.* **46**, 431-449 (1988).

[10] W. A. DAY: Existence of a property of solutions of the heat equation subject to linear thermoelasticity and other theories. *Quart. Appl. Math.* **40**, 319-330, (1982).

[11] W. A. DAY: A decreasing property of solutions of a parabolic equation with applications to thermoelasticity and other theories, *Quart. Appl. Math.* Vol. XLIV, (1983), 468-475.

[12] A. FRIEDMAN: Monotonic decay of solutions of parabolic equation with nonlocal boudary conditions, *Quart. Appl. Math.* Vol. XLIX, (1986), 468-475.

[13] C. F. GERALD: *Applied Numerical Analysis*, (fourth edition) Addison-Wesley, California, (1989).

[14] A. B. GUMEL, W. T. ANG and E. H. TWIZELL: Efficient algorithms for the two-dimensional diffusion equation subject to specification of mass , *Intern. J. Computer Math.*, **64**, 153-163, (1997).

[15] A. R. GOURLAY and J. Li MORRIS: The extrapolation of first-order methods for parabolic partial differential equations, II, SIAM, J. Numer. Anal. **17**, 641-655, (1980).

[16] J. D. LAMBERT: Numerical methods for ordinary differential systems: The Initial-Value Problem, John Wiley and Sons, Chichester, (1991).

[17] J. D. LAWSON and J. Li MORRIS: The extrapolation of first-order methods for parabolic partial differential equations, I, *SIAM, J. Numer. Anal.*, **15**, 1212-1224, (1978).

[18] E. H. TWIZELL, A. B. GUMEL and M. A. ARIGU: Second-orde L_0-stable methods for the heat equation with time-dependent boundary conditions, *Adv. Comput. Math.*, **6**, 333-352, (1996).

[19] D. A. VOSS and A. Q .M. KHALIQ : Time splitting algorithms for semidiscretized linear parabolic PDEs based on rational approximants with distinct poles, *Adv. Comput. Math.*, **6**, 353-363, (1996).

JOURNAL OF CONCRETE AND APPLICABLE MATHEMATICS,VOL.3,NO.4,469-480,
2005,COPYRIGHT 2005 EUDOXUS PRESS,LLC

Random Fixed Points and Iteration Process for Asymptotically Nonexpansive Random Maps

Ismat Beg

Centre for Advanced Studies in Mathematics and

Department of Mathematics,

Lahore University of Management Sciences,

54792 Lahore, PAKISTAN.

E-mail: ibeg@lums.edu.pk

Abstract— The purpose of this article is to prove existence of random fixed point for asymptotically nonexpansive random maps. We also construct a Mann type iteration process for asymptotically nonexpansive random maps which converges to the random fixed point.

Key words and phrases. Random fixed point, asymptotically nonexpansive map, Banach space.

2000 Mathematics Subject Classifications. 47H40, 47H10, 47H09.

1. INTRODUCTION

In rapidly developing area of nonlinear theory of differential equations, many important results have been obtained by the use of nonlinear functional analysis based on fixed point theory. Over the last thirty years, random operator theory has grown into a full fledged research area. The various ideas associated with random fixed point theory can be used to form a particularly elegant approach for the solution of nonlinear random systems (see [4]). Random fixed point theorems for random contraction mappings on Polish spaces were first proved by Spacek [21] and Hans [9,10]. Subsequently Bharucha-Reid

[5] (see also Mukherjea [16]) have given sufficient conditions for a stochastic analogue of Schauder's fixed point theorem for a random operator. Itoh [12,13] introduced random condensing operators and considerably improved their results. Sehgal and Water [19,20] considered Browder-Fan type random operators and as a consequence obtained a stochastic generalization of the well known Rothe fixed point theorem. Recently Papageorgiou [17], Lin [14], Xu [23], Beg [1], Tan and Yuan [22], Beg and Shahzad [2,3] and many other authors have studied the fixed points of random maps. The aim of this note is to study in continuation the fixed points of asymptotically nonexpansive random maps (i.e. ill conditioned random problems) and to construct a Mann type iteration scheme [15] which converges to a fixed point of the aysmptotically nonexpansive random map.

2. PRELIMINARIES

A Banach space X is called *uniformly convex* if for each $\varepsilon > 0$ there is a function $\delta : (0, 2] \to (0, 1]$ such that if $\|x\| \leq d$, $\|y\| \leq d$, $\|x - y\| \geq \varepsilon$ then $\|\frac{x+y}{2}\| \leq (1-\delta(\frac{\varepsilon}{d}))d$. Moreover δ is increasing. In the sequel all normed spaces are assumed to be taken over the real.

Let (Ω, Σ) be a measurable space (Σ = sigma algebra) and E a nonempty subset of a Banach space X. A mapping $\xi : \Omega \to X$ is *measurable* if and only if $\xi^{-1}(U) \in \Sigma$ for each open subset U of X. The mapping $T : \Omega \times E \to X$ is a *random map* if and only if for each fixed $x \in E$, the mapping $T(., x) : \Omega \to X$ is measurable, and it is *completely continuous* if for each $\omega \in \Omega$, the mapping $T(\omega, .) : E \to X$ is completely continuous. A measurable mapping $\xi : \Omega \to X$ is a *random fixed point* of the random map $T : \Omega \times E \to X$ if and only if $T(\omega, \xi(\omega)) = \xi(\omega)$ for each $\omega \in \Omega$. For a subset E of X, let $diam(E)$ denote the diameter, $cl(E)$ the closure, $co(E)$ the convex hull, $clco(E)$ the closed convex hull of the set E, $S(x_0, r)$ the spherical ball centred at x_0 with radius r i.e. $S(x_0, r) = \{x \in X : \|x_0 - x\| \leq r\}$ and $T^n(\omega, x)$ the nth iterate $T(\omega, T(w, T(w, ...T(\omega, x)...)))$.

Definition 2.1. Let E be a subset of a Banach space X and $T : \Omega \times E \to E$ be a random map.

(i). The random map T is said to be *nonexpansive* if for arbitrary $x, y \in E$,

$\|T(\omega, x) - T(\omega, y)\| \leq \|x - y\|$,

for each $\omega \in \Omega$.

(ii).The random map T is called *asymptotically nonexpansive* if for each $x, y \in E$,

$\|T^n(\omega, x) - T^n(\omega, y)\| \leq k_n\|x - y\|$,

for each $\omega \in \Omega$ and where $\{k_n\}$ is a decreasing sequence of real numbers such that $\lim_{n\to\infty} k_n = 1$.

(iii). The random map T is said to be *uniformly L-Lipschitzian* $(L > 0)$ if and only if for all $x, y \in E$,

$\|T^n(\omega, x) - T^n(\omega, y)\| \leq L\|x - y\|$

for each $\omega \in \Omega$.

Theorem 2.2. [24,Theorem 2]. Let $p > 1$, $r > 0$ be two fixed real numbers. Then the Banach space X is uniformly convex if and only if there exists a continuous, strictly increasing and convex function $g : \mathbf{R}^+ \to \mathbf{R}^+$, $g(0) = 0$, such that,

$$\|\alpha x + (1 - \alpha)y\|^p \leq \alpha\|x\|^p + (1 - \alpha)\|y\|^p - W_p(\alpha)g(\|x - y\|), \qquad (1)$$

for all x, y in $S(\mathbf{0}, r)$, $\alpha \in [0, 1]$ where $W_p(\alpha) = \alpha(1 - \alpha)^p + \alpha^p(1 - \alpha)$.

For more details and other related results we refer to [2,5,7,24].

3. THE RESULTS

Our principal result in this section is the following random version of a generalization due to Goebel-Kirk [8] of Browder [6] fixed point theorem.

Theorem 3.1. Let E be a nonempty closed bounded and convex subset of a separable uniformly convex Banach space X. Let $T : \Omega \times E \to E$ be an asymptotically nonexpansive random map. Then T has a random fixed point.

Proof. Let $\xi : \Omega \to E$ be a measurable map. Let $w \in \Omega$ be fixed, and let the set $R_{\xi(w)}$ consists of those real numbers r for which there exists an integer

k such that

$$E \cap (\cap_{n=k}^{\infty} S(T^n(w, \xi(w)), r) \neq \phi.$$

Since diam $(E) \in R_{\xi(w)}$, therefore $R_{\xi(w)} \neq \phi$. Let $r_0 = \inf R_{\xi(w)}$. For each $\varepsilon > 0$,

$$(\cup_{k=1}^{\infty} (\cap_{n=k}^{\infty} S(T^n(w, \xi(w)),\ r_0 + \varepsilon))) \cap E,$$

are nonempty and convex. The reflexivity of X further implies that

$$\cap_{\varepsilon>0} (c\ell (\cup_{k=1}^{\infty} (\cap_{n=k}^{\infty} S(T^n(w, \xi(w)), r_0 + \varepsilon)))) \cap E \neq \phi.$$

Let $\eta : \Omega \to E$ be a measurable map so that for each $w \in \Omega$,

$$\eta(w) \in \cap_{\varepsilon>0} (c\ell (\cup_{k=1}^{\infty} (\cap_{n=k}^{\infty} S(T^n(w, \xi(w)), r_0 + \varepsilon)))) \cap E,$$

(for existence of η see [11]).

Let $\alpha > 0$, then there exists an integer n_0 such that if $n \geq n_0$, then

$$\|\eta(w) - T^n(w, \xi(w))\| \leq r_0 + \alpha.$$

Suppose that the sequence $\{T^n(w, \eta(w))\}$ does not converge to $\eta(w)$ (i.e. η is not a random fixed point of T).

Then there exist $\varepsilon > 0$ and a subsequence $\{T^{n_i}(w, \eta(w))\}$ of $\{T^n(w, \eta(w))\}$ such that

$$\|T^{n_i}(w, \eta(w)) - \eta(w)\| \geq \varepsilon,\ i = 1, 2, 3, \cdots .$$

For $m > n$,

$$\|T^n(w, \eta(w)) - T^m(w, \eta(w))\| \leq k_n \|\eta(w) - T^{m-n}(w, \eta(w))\|,$$

where k_n is the Lipschitz constant for T^n obtained from the definition of asymptotic nonexpansiveness. Suppose $r_0 > 0$ and choose $\rho > 0$ so that

$$\left(1 - \delta(\frac{\varepsilon}{r_0 + \rho})\right)(r_0 + \rho) < r_0.$$

Select p so that $\|\eta(w) - T^p(w, \eta(w))\| \geq \varepsilon$, and also so that

$$k_p(r_0 + \frac{\rho}{2}) \leq r_0 + \rho.$$

If $n_0 \geq p$ is sufficiently large, then $m > n_0$ implies that

$$\|\eta(w) - T^{m-p}(w, \xi(w))\| \leq r_0 + \frac{\rho}{2}.$$

It further implies that

$$\|T^p(w, \eta(w)) - T^m(w, \xi(w))\| \leq k_p \|\eta(w) - T^{m-p}(w, \xi(w))\| \leq r_0 + \rho,$$

also

$$\|\eta(w) - T^m(w, \xi(w))\| \leq r_0 + \rho.$$

Thus by uniform convexity of X, if $m > n_0$, then

$$\|\frac{\eta(w) + T^p(w, \eta(w))}{2} - T^m(w, \xi(w))\| \leq \left(1 - \delta\left(\frac{\varepsilon}{r_0 + \rho}\right)\right)(r_0 + \rho) \leq r_0,$$

and this contradicts the definition of r_0. Hence we conclude $r_0 = 0$ or $T(w, \eta(w)) = \eta(w)$ for each $w \in \Omega$. But $r_0 = 0$ implies $\{T^n(w, \xi(w))\}$ is a Cauchy sequence yielding $\lim_{n\to\infty} T^n(w, \xi(w)) = \eta(w) = T(w, \eta(w)$, for each $w \in \Omega$. Therefore for each $w \in \Omega$,

$$\cap_{\varepsilon>0} \left(c\ell \left(\cup_{k=1}^{\infty} \left(\cap_{n=k}^{\infty} S\left(T^n(w, \xi(w)),\ r_0 + \varepsilon\right)\right)\right)\right) \cap E = \{\eta(w)\}.$$

Hence η is a random fixed point of T. This completes the proof.

Let X be a separable normed space, E be a nonempty convex subset of X and $L > 0$. Let $T : \Omega \times E \to E$ be a uniformly L-Lipschitzian random map. We define a Mann [15] type random iteration process as follows.

Let $\xi_0 : \Omega \to E$ be any measurable map. Then the following measurable maps are iteratively defined as follows:

$$\eta_n(w) = \beta_n T^n(w, \xi_n(w)) + (1 - \beta_n)\xi_n(w), \tag{2}$$

and

$$\xi_{n+1}(w) = \alpha_n T^n(w, \eta_n(w)) + (1 - \alpha_n)\xi_n(w) \tag{3}$$

for each $w \in \Omega$ and $n = 0, 1, 2, 3, \cdots$ where $\{\alpha_n\}$ and $\{\beta_n\}$ satisfy $\varepsilon \leq 1 - \alpha_n, 1 - \beta_n \leq 1 - \varepsilon$ for all n and some $\varepsilon > 0$. Since E is nonempty convex subset

of X therefore the sequences of measurable functions $\{\xi_n\}$ and $\{\eta_n\}$ are well defined.

Lemma 3.2. Let $T : \Omega \times E \to E$ be a uniformly L-Lipschitzian random map. Then for all $n = 0, 1, 2, \cdots$.

$$\|\xi_n(w) - T(w, \xi_n(w))\| \leq \|T^n(w, \xi_n(w)) - \xi_n(w)\|$$

$$+ \left(L + 3L^2 + 2L^3\right) \|T^{n-1}(w, \xi_{n-1}(w)) - \xi_{n-1}(w)\|. \tag{4}$$

Proof. For all $n = 0, 1, 2, \cdots$, we have

$$\|\xi_n(w) - T(w, \xi_n(w))\| \leq \|\xi_n(w) - T^n(w, \xi_n(w))\|$$

$$+\|T^n(w, \xi_n(w)) - T(w, \xi_n(w))\|. \tag{5}$$

Now,

$$\begin{aligned} \|T^n(w, \xi_n(w)) &- T(w, \xi_n(w))\| = \|T(w, T^{n-1}(w, \xi_n(w)) - T(w, \xi_n(w))\| \\ &\leq L\|T^{n-1}(w, \xi_n(w)) - \xi_n(w)\| \end{aligned}$$

$$\leq L\|T^{n-1}(w, \xi_n(w)) - \left\{\alpha_{n-1}T^{n-1}(w, \eta_{n-1}(w)) + (1 - \alpha_{n-1})\xi_{n-1}(w)\right\}\|$$

$$\leq L\left[\alpha_{n-1}\|T^{n-1}(w, \xi_n(w)) - T^{n-1}(w, \eta_{n-1}(w))\|\right.$$

$$\left.+(1 - \alpha_{n-1})\|T^{n-1}(w, \xi_n(w)) - \xi_{n-1}(w)\|\right]$$

$$\leq L\left[L\|\xi_n(w) - \eta_{n-1}(w)\| + \|T^{n-1}(w, \xi_n(w)) - \xi_{n-1}(w)\|\right]. \tag{6}$$

Also,

$$\|\xi_n(w) - \eta_{n-1}(w)\| \leq \|\xi_n(w) - \xi_{n-1}(w)\| + \|\xi_{n-1}(w) - \eta_{n-1}(w)\| \tag{7}$$

More over,

$$\|\xi_{n-1}(w) - \eta_{n-1}(w)\| = \|\xi_{n-1}(w) - \{\beta_{n-1}T^{n-1}(w, \xi_{n-1}(w))$$

$$+(1 - \beta_{n-1})\xi_{n-1}(w)\}\|$$

$$\leq \|T^{n-1}(w, \xi_{n-1}(w)) - \xi_{n-1}(w)\|. \tag{8}$$

And,

$$\|\xi_n(w) - \xi_{n-1}(w)\| = \|\alpha_{n-1}T^{n-1}(w, \eta_{n-1}(w)) + (1 - \alpha_{n-1})\xi_{n-1}(w) - \xi_{n-1}(w)\|$$

$$= \alpha_{n-1}\|T^{n-1}(w, \eta_{n-1}(w)) - \xi_{n-1}(w)\|$$

$$\leq \|\xi_{n-1}(w) - T^{n-1}(w, \xi_{n-1}(w))\|$$

$$+\|T^{n-1}(w, \xi_{n-1}(w)) - T^{n-1}(w, \eta_{n-1}(w))\|$$

$$\leq \|\xi_{n-1}(w) - T^{n-1}(w, \xi_{n-1}(w))\| + L\|\xi_{n-1}(w) - \eta_{n-1}(w)\|$$

$$\leq \|\xi_{n-1}(w) - T^{n-1}(w, \xi_{n-1}(w))\|$$

$$+L\|\xi_{n-1}(w) - \left\{\beta_{n-1}T^{n-1}(w, \xi_{n-1}(w)) + (1 - \beta_{n-1})\xi_{n-1}(w)\right\}\|$$

$$\leq (1 + L)\|\xi_{n-1}(w) - T^{n-1}(w, \xi_{n-1}(w))\|. \tag{9}$$

In equalities (7), (8) and (9) imply,

$$\|\xi_n(w) - \eta_{n-1}(w)\| \leq (2 + L)\|\xi_{n-1}(w) - T^{n-1}(w, \xi_{n-1}(w))\|. \tag{10}$$

Also note that,

$$\|T^{n-1}(w, \xi_n(w)) - \xi_{n-1}(w)\|$$

$$\leq \|T^{n-1}(w, \xi_n(w)) - T^{n-1}(w, \xi_{n-1}(w))\|$$

$$+\|T^{n-1}(w, \xi_{n-1}(w)) - \xi_{n-1}(w)\|$$

$$\leq L\|\xi_n(w) - \xi_{n-1}(w)\| + \|T^{n-1}(w, \xi_{n-1}(w)) - \xi_{n-1}(w)\|.$$

(by (9))

$$\leq \{1 + L(1 + L)\}\|T^{n-1}(w, \xi_{n-1}(w)) - \xi_{n-1}(w)\| \tag{11}$$

Inequalities (6), (10) and (11) imply,

$$\|T^n(w, \xi_n(w)) - T(w, \xi_n(w))\|$$

$$\leq L\left[L(2+L)\|T^{n-1}(w,\xi_{n-1}(w))-\xi_{n-1}(w)\|\right.$$
$$\left.+\{1+L(1+L)\}\|T^{n-1}(w,\xi_{n-1}(w))-\xi_{n-1}(w)\|\right]$$

$$=(L+3L^2+2L^3)\|T^{n-1}(w,\xi_{n-1}(w))-\xi_{n-1}(w)\| \tag{12}$$

Inequalities (5) and (12) yield the required inequality (2). This completes the proof.

Theorem 3.3. Let X be a separable uniformly convex Banach space; $\phi \neq E \subset X$, and E closed bounded and convex. Let $T : \Omega \times E \to E$ be a asymptotically nonexpansive random map with $\{k_n\}$ satisfying

$$k_n \geq 1,\ \sum_{n=1}^{\infty}\left(k_n^2-1\right)<\infty,\ \varepsilon \leq 1-\alpha_n \leq 1-\varepsilon$$

for all n and some $\varepsilon > 0$.

Let $\xi_0 : \Omega \to E$ be a measurable map and define

$$\xi_{n+1}(w)=\alpha_n T^n(w,\xi_n(w))+(1-\alpha_n)\xi_n(w)$$

for $n \geq 0$. Then

$$\lim_{n\to\infty}\|\xi_n(w)-T(w,\xi_n(w)\|=0.$$

Proof. Set $M = [diam(E)]^2$. From Theorem 3.1, T has a random fixed point ξ (say) (i.e. $\xi : \Omega \to E$ a measurable map and $T(w,\xi(w)) = \xi(w)$). Using inequality (1), with $r > 1$ such that $E \subseteq S(\mathbf{0}, r)$, we have,

$$\|\xi_{n+1}(w)-\xi(w)\|^p=\|\alpha_n T^n(w,\xi_n(w))+(1-\alpha_n)\xi_n(w)-\xi(w)\|^p$$

$$\leq \alpha_n\|T^n(w,\xi_n(w))-\xi(w)\|^p+(1-\alpha_n)\|\xi_n(w)-\xi(w)\|^p$$
$$-W_p(\alpha_n)g\left(\|\xi_n(w)-T^n(w,\xi_n(w))\|\right)$$

$$\leq \alpha_n k_n^p\|\xi_n(w)-\xi(w)\|^p+(1-\alpha_n)\|\xi_n(w)-\xi(w)\|^p$$
$$-W_p(\alpha_n)g\left(\|\xi_n(w)-T^n(w,\xi_n(w))\|\right)$$

$$\leq \|\xi_n(w)-\xi(w)\|^p+\left(\alpha_n k_n^p-\alpha_n\right)M$$

$$-W_p(\alpha_n)g\left(\|\xi_n(w) - T^n(w, \xi_n(w))\|\right) \tag{13}$$

Since $W_p(\alpha_n) \geq 2\varepsilon^{p+1}$ (See Xu [24]), thus from (13), we obtain

$$0 \leq 2\varepsilon^{p+1} g\left(\|\xi_n(w) - T^n(w, \xi_n(w))\|\right)$$

$$\leq \|\xi_n(w) - \xi(w)\|^p - \|\xi_{n+1}(w) - \xi(w)\|^p - \alpha_n M\left(k_n^p - 1\right).$$

It implies that

$$2\varepsilon^{p+1} \sum_{n=0}^{m} g\left(\|\xi_n(w) - T^n(w, \xi_n(w))\|\right)$$

$$\leq \|\xi_0(w) - \xi(w)\|^p + (1-\varepsilon)M \sum_{n=0}^{m}(k_n^p - 1) \tag{14}$$

¿From hypothesis on $\{\xi_n(w)\}$, inequality (14) implies that the series on the left converges. Therefore

$$\lim_{n\to\infty} g\left(\|\xi_n(w) - T^n(w, \xi_n(w))\|\right) = 0. \tag{15}$$

Since g is continuous at 0 and strictly increasing therefore (15) implies that

$$\lim_{n\to\infty} \|\xi_n(w) - T^n(w, \xi_n(w))\| = 0.$$

Lemma 3.2. now further implies

$$\lim_{n\to\infty} \|\xi_n(w) - T(w, \xi_n(w))\| = 0.$$

This completes the proof.

Now we are able to show that $\{\xi_n\}$ converges to the random fixed point of $T : \Omega \times E \to E$.

Theorem 3.4. Let E be a nonempty closed bounded and convex subset of a separable uniformly convex Banach space X. Let $T : \Omega \times E \to E$ be a completely continuous asymptotically nonexpansive random map with $\{k_n\}$ satisfying $k_n \geq 1$ and $\sum_{n=1}^{\infty}\left(k_n^p - 1\right) < \infty$; $r = \max\{2, p\}$; $\varepsilon \leq \alpha_n \leq 1 - \varepsilon$ for all n and some $\varepsilon > 0$. Choose $\xi_0 : \Omega \to E$ a measurable map, and define

$\xi_{n+1}(w) = \alpha_n T^n(w, \xi_n(w)) + (1-\alpha_n)\xi_n(w),\ n \geq 0$. Then $\{\xi_n\}$ converges to some random fixed point ξ of T.

Proof. Clearly, $clco(\{\xi_0(w)\} \cup T(\Omega \times E))$ is a compact subset of E containing $\{\xi_n(w)\}$. Therefore there exists a measurable map $\xi : \Omega \to E$ and some subsequence $\{\xi_{n_i}(w)\}$ of $\{\xi_n(w)\}$ such that $\lim_{i\to\infty} \xi_{n_i}(w) = \xi(w)$. But T is continuous and $\lim_{n\to\infty} \|\xi_n(w) - T(w, \xi_n(w)) = 0$ by Theorem 3.3. Therefore $T(w, \xi(w)) = \xi(w)$. It implies that

$$\begin{aligned}
\|\xi_{n+1}(w) - \xi(w)\| &= \|\alpha_n T^n(w, \xi_n(w)) + (1-\alpha_n)\xi_n(w) - \xi(w)\| \\
&\leq \alpha_n \|T^n(w, \xi_n(w)) - \xi(w)\| + (1-\alpha_n)\|\xi_n(w) - \xi(w)\| \\
&= \alpha_n \|T^n(w, \xi_n(w)) - T(w, \xi(w))\| + (1-\alpha_n)\|\xi_n(w) - \xi(w)\| \\
&\leq (\alpha_n k_n + (1-\alpha_n))\, \|\xi_n(w) - \xi(w)\|
\end{aligned}$$

for all $n \geq 0$. Since $\Pi_{n=1}^{\infty} k_n$ converges and $\lim_{i\to\infty} \xi_{n_i}(w) = \xi(w)$. Therefore $\lim_{n\to\infty} \xi_n(w) = \xi(w)$. Hence $\{\xi_n\}$ converges to the random fixed point ξ of T.

Remark 3.5. Schu [18, Theorem 1.5] is a particular case of the deterministic analogue of our Theorem 3.4.

References

[1] I. Beg, Random fixed points of random operators satisfying semicontractivity conditions, Math. Japon., 46(1)(1997), 151-155.

[2] I. Beg and N. Shahzad, Random fixed point theorems for nonexpansive and contractive type random operators on Banach spaces, J. Appl. Math. Stochas. Anal., 7(4)(1994), 569 - 580.

[3] I. Beg and N. Shahzad, Some random approximation theorem with applications, Nonlinear Analysis, 35(1999), 609-616.

[4] A. T. Bharucha-Reid, Random Integral Equations, Academic Press, New York , 1972.

[5] A. T. Bharucha-Reid, Fixed point theorems in probabilistic analysis, Bull. Amer. Math. Soc. , 82(1976), 641 - 657.

[6] F. E. Browder, Nonexpansive nonlinear operators in Banach spaces, Proc. Natl. Acad. Sci. U.S.A., 54 (1965), 1041-1044.

[7] D. L. Cohn, Measure Theory, Birkhauser, Boston - 1980.

[8] K. Goebel and W. A. Kirk, A fixed point theorem for asymptotically nonexpansive mappings, Proc. Amer. Math. Soc.,35(1)(1972), 171-174.

[9] O. Hans, Reduzierende zufällige transformationen, Czechoslovak Math. Jour., 7(1957), 154 -158.

[10] O. Hans, Random operator equations, In: Proc. 4th Berkeley Symposium on Mathematical Statistics and Probability Vol. II, Part I, 185-202, University of California Press, Berkeley 1961.

[11] C. J. Himmelberg, Measurable relations, Fund. Math., 87(1975), 53 - 72.

[12] S. Itoh, A random fixed point theorem for a multivalued contraction mapping, Pacific J. Math.,68(1977), 85-90.

[13] S. Itoh, Random fixed point theorems with an application to random differential equations in Banach spaces, J. Math. Anal. Appl., 67(1979), 261-273.

[14] T. C. Lin, Random approximations and random fixed point theorems for nonself maps, Proc. Amer. Math. Soc., 103(1988), 1129 - 1135.

[15] W. R. Mann, Mean valued methods in iteration, Proc. Amer. Math. Soc., 4(1953), 506-510.

[16] A. Mukherjee, Random transformations of Banach spaces, Ph.D. Dissertation, Wayne State Univ. Detroit, Michigan -1968.

[17] N. S. Papageorgiou, Random fixed point theorems for measurable multifunctions in Banach spaces, Proc. Amer. Math. Soc., 97(1986), 507 - 514.

[18] J. Schu, Iterative construction of fixed points of asymptotically nonexpansive mappings, J. Math. Anal. Appl., 158(1991), 407 - 413.

[19] V. M. Sehgal and C. Waters, Some random fixed point theorems, Contemporary Math., 21(1983), 215 - 218.

[20] V. M. Sehgal and C. Waters, Some random fixed point theorems for condensing operators, Proc. Amer. Math. Soc., 90(3)(1984),425-429.

[21] A. Spacek, Zufällige gleichungen, Czechoslovak Math. Jour. 5(1955), 462-466.

[22] K. K. Tan and X. Z. Yuan, Random fixed point theorems and approximations, Stochas. Anal. Appl., 15(1)(1997), 103 - 123.

[23] H. K. Xu, Some random fixed point theorems for condensing and non expansive operators, Proc. Amer. Math. Soc., 110(1990), 395 - 400.

[24] H. K. Xu, Inequalities in Banach spaces with applications, Nonlinear Analysis, 16(1991), 1127-1138.

JOURNAL OF CONCRETE AND APPLICABLE MATHEMATICS,VOL.3,NO.4,481-499,
2005,COPYRIGHT 2005 EUDOXUS PRESS,LLC

A note on a Liapounov-like theorem for some finitely additive measures and applications

Anna Martellotti – Anna Rita Sambucini

Dipartimento di Matematica e Informatica
1, Via Vanvitelli - 06123-I, Perugia (ITALY)
e-mail: amart@dipmat.unipg.it, matears1@unipg.it

Abstract

We give some results about the convexity of a pair of finitely additive measures and we apply them to derive the convexity of the Aumann integral of a suitable multifunction.

1991 AMS Mathematics Subject Classification: **28B20, 26E25, 46B20, 54C60**
Key words: Aumann integral, finitely additive measures, selections, convex measures, Stone isomorphism.

1. Introduction

The classical core-Walras equivalence result [13, Theorem II-2.1] is stated for a finite-dimensional commodity space, and a space of agents (Ω, Σ, μ) represented by a non-atomic, positive, countably additive measure space. This celebrated result of Equilibrium Theory has been extended in several directions: in particular ([1, 3, 28]) discuss and face the problem of extending it to the case of *strongly non-atomic* finitely additive measures. The proof of the classical result is based upon an idea of Aumann, and makes use of the geometrical and topological properties of the multivalued integral of a suitable

multifunction ranging on the commodity space.

The Aumann integral of a Banach-valued multifunction with respect to a finitely additive measure has been considered in [21] and [22]: in [21] the investigation concerned integrands with compact and convex values, while [22] examined the case of integrands with weakly compact and convex values; (see also [26] for a survey on this topic).

Recently, in [23] we have examined the case of multifunctions of the type

$$F(\omega) = [\Gamma(\omega) - e(\omega)] \cup \{0\}, \tag{1}$$

where Γ is a simple multifunction with values in the hyperspace of closed and convex values of a Banach space X, $e \in L^1_\mu(X)$ and μ is non atomic and countably additive.

Integrands of this form are those that occur in the core-Walras equivalence: the assumption on Γ has a meaningful interpretation from the point of view of the economic model.

The properties of the multivalued integral are usually achieved applying the classical Liapounov Theorem, together with its infinite dimensional versions. It is well known that Liapounov Theorem does not extend to finitely additive measures: weakened forms of it in this more general setting have been obtained by several authors [18, 6, 7, 29, 8, 2, 3]. A survey of the most important results can be found in [19].

What we obtained in [23] was that, if e has Liapounov indefinite integral, then F has convex Aumann integral. The proof of the result was based upon the following result

Theorem A ([23] Theorem 3.7) *Let X and Y be two Banach spaces, with X satisfying the* (RNP), *$\mu : \Sigma \to \mathbb{R}^+_0$ a non-atomic countably additive measure, $f = \sum_{i=1}^{n} x_i 1_{E_i}$ a Y-valued, simple function, and $n_2 = \int_\cdot e d\mu$ an X-valued Liapounov measure. Then, setting $n_1 = \int_\cdot f d\mu$, the range of the pair (n_1, n_2) is convex and compact in $Y \times X_w$.*

This paper is concerned with the Aumann integral for integrands of the form (1), but when μ is simply finitely additive: in this setting Theorem A above is in general false; a counterexample can be obtained by means of the results in [8].

However we shall show, through a completely different path, that the convexity of the Aumann integral of F can be reobtained also in the finitely additive setting, when the commodity space X is a Banach lattice and the indefinite integral of e satisfies suitable assumptions.

2. Preliminaries and definitions

Throughout this paper X will be a reflexive, separable Banach lattice, X^+ its positive cone. With X^* we denote the topological dual and with $X_1,\ X_1^*$ the unit balls of X and X^* respectively. We denote by X_w the space X equipped with its weak topology.
Let Ω be a set, Σ a σ-algebra of subsets of Ω and $\mu : \Sigma \to [0, +\infty[$ a finitely additive bounded measure. In accordance to [11, Chapter III] we denote by $L^1_\mu(X)$ the space of X-valued, μ-integrable functions f. When $X = \mathbb{R}$ we shall simply write L^1_μ. Throught the paper we will use the symbol μ to denote a scalar measure, while with the symbol m we denote a vector valued one.

Definition 1 A finitely additive vector measure $m : \Sigma \to X$ is called *Liapounov* if, for every $E \in \Sigma$, $\quad m(\Sigma_E) := \{m(A), A \in \Sigma \cap E\}$ is convex and weakly compact for every $E \in \Sigma$. Since we have assumed that X is a reflexive Banach space it is enough to assume that $m(\Sigma_E)$ is bounded, closed and convex for every $E \in \Sigma$. If, for every $E \in \Sigma$, $m(\Sigma_E)$ is only convex, we will say that m is a *convex measure.* If, for every $E \in \Sigma$, there exists $B \in \Sigma_E$ such that $m(B) = \frac{1}{2}m(E)$, we will say that m is a *semiconvex measure.*

We remind that for a scalar finitely additive, bounded measure $\mu : \Sigma \to [0, +\infty)$ the conceps of strong continuity, semiconvexity and Liapounov are equivalent [8, 18]; where μ is *strongly continuous* if for every $\varepsilon > 0$ there exists a finite decomposition of Ω, $A_i \in \Sigma,\ i = 1, \ldots, n$ such that $\mu(A_i) \leq \varepsilon$.

For a finitely additive vector measure we will denote by $|m|$ the variation of m, defined for every $E \in \Sigma$, by:

$$|m|(E) = \sup_{\Pi} \sum_{A_i \in \Pi} \|m(A_i)\|$$

where the supremum is taken over all the finite decompositions Π of the set E. Moreover we denote by $\|m\|$ its semivariation, given by

$$\|m\|(E) = \sup\{|x^* m|(E),\ x^* \in X^*,\ \|x^*\| \leq 1\}.$$

It is known that:

$$\sup\{\|m(A)\|, A \in \Sigma \cap E\} \leq \|m\|(E) \leq 4 \sup\{\|m(A)\|, A \in \Sigma \cap E\},$$

see for example [9, Proposition 1.1].

In the framework of [27] we will make use of the Stone extension; more precisely $\mathcal{G}$ will be the Stone algebra associated to Σ and $\tau : \Sigma \to \mathcal{G}$ the Stone isomorphism.
With $\overline{\mu} : \mathcal{G}_\sigma \to [0, +\infty[$ we will denote the extended measure of μ, where $\mathcal{G}_\sigma$ is the σ-algebra generated by $\mathcal{G}$. Observe that if μ is strongly continuous then its Stone extension $\overline{\mu}$ is non atomic and therefore Liapounov. Moreover, by [12], if $f \in L^1_\mu(X)$ then it is possible to define its Stone extension as a map $\overline{f} \in L^1_{\overline{\mu}}(X)$ such that for every $E \in \Sigma$,

$$\int_E f d\mu = \int_{\tau(E)} \overline{f} d\overline{\mu} \tag{2}$$

where the left hand side is defined in accordance to [11, Chapter III].
As a consequence if $f \in L^1_\mu(X)$ has Liapounov indefinite integral, its Stone extension $\overline{f}$ has the same property. It was also showed in [15] that $\|\overline{f}\| = \overline{\|f\|}$ $\overline{\mu}$-almost everywhere.

Let $m : \Sigma \to X$ be a vector-valued finitely additive measure. We say that m is *s-bounded* if $\lim_{n\to\infty} m(A_n) = 0$ for every sequence $(A_n)_n$ of pairwise disjoint sets in Σ.

Definition 2 A positive finitely additive measure $\sigma : \Sigma \to [0, \infty[$ is a *control* for m if and only if $\|m\| \sim \sigma$, in the sense that for every $\varepsilon > 0$ there exists $\delta > 0$ such that the following implications hold:

- if $\sigma(A) < \delta$ then $\|m\|(A) < \varepsilon$;
- if $\|m\|(A) < \delta$ then $\sigma(A) < \varepsilon$.

A control σ is said to be a *Rybakov control* if there exists a functional $x^* \in X^*$ such that $\sigma = |x^*m|$.

Remark 1 In [10, 25] the following equivalences were proved: a finitely additive measure m is s-bounded if and only if there exists a control for m if and only if there exists a Rybakov control for m. If m is also of bounded variation then its variation is equivalent to a Rybakov control for m. The following proof of this equivalence was communicated to us by one of the referees. If $\sigma = |x^*m|$ then obvioulsy $\sigma \ll |m|$. To prove the converse, for every $F = \{x_1^*, \dots, x_n^*\}$ in X_1^* let $\eta_F := \bigvee_{i=1}^n |x_i^*m|$ be the lattice supremum of the finitely additive measures $|x_1^*m|, \dots, |x_n^*m|$. Since all the η_F are dominated by $|m|$, then they are uniformly s-bounded and therefore they are uniformly σ-continuous. Hence the set-wise supremum $|m|(E) = \sup\{\eta_F(E), F \subset X_1^*, F \text{ finite } \}$ is σ-continuous. So, in this case, $\|m\| \sim |m|$. Moreover if m and σ are countably additive then the $\varepsilon - \delta$ absolute continuity is equivalent to $0 - 0$ absolute continuity.

3. A Liapounov result

We will now show that if m is X^+-valued and s-bounded then it admits a Rybakov control of the form $\sigma = y^*m$ for some $y^* \in (X^*)^+$.

Lemma 1 *If $m : \Sigma \to X^+$ is a s-bounded finitely additive measure then there exists $y^* \in (X^*)^+$ such that $\|m\| \sim y^*m$.*

Proof: Let $\sigma = |x_0^*m|$ be a Rybakov control for m, and let $y^* = |x_0^*|$ in the Banach lattice X^*.
Then, easily, $|x_0^*(x)| \leq y^*(x)$ for every $x \in X^+$. Let now $\varepsilon > 0$ be fixed and consider δ according to the absolute continuity of $\|m\|$ with respect to σ; let $A \in \Sigma$ be such that $y^*m(A) \leq \delta$; we want to show that $\|m\|(A) \leq \varepsilon$. Indeed let Π be an arbitrary finite partition of A, since m is X^+-valued we have

$$\sum_{B \in \Pi} |x_0^*m(B)| \leq \sum_{B \in \Pi} y^*m(B) = y^*m(A) \leq \delta.$$

Taking the supremum with respect to Π we have $\sigma(A) \leq \delta$, which in turn yields $\|m\|(A) \leq \varepsilon$.
Conversely we prove now that $y^*m \ll \|m\|$. Let $\varepsilon > 0$ be fixed and consider

$\delta = \varepsilon\|y^*\|^{-1}$. If $\|m\|(A) < \delta$ then $y^*m(A) \leq \|y^*\|\ \|m\|(A) \leq \|y^*\|\ \delta < \varepsilon$. □

We will show now that, also in the finitely additive case, if m is of bounded variation it is possible to find a control which is equivalent to the variation of m.

Proposition 1 *If $m : \Sigma \to X^+$ is a finitely additive measure with bounded variation, then there exists $y^* \in (X^*)^+$ such that $|m| \sim y^*m$.*

Proof: By Lemma 1 there exists $y^* \in (X^*)^+$ such that y^*m is a Rybakov control, namely $y^*m \sim \|m\|$. Since m is of bounded variation then, by Remark 1, $\|m\|$ is equivalent to $|m|$. This concludes the proof. □

Proposition 2 *Let $m : \Sigma \to X^+$ be an s-bounded finitely additive measure. The following are equivalent:*

2.1 *m is semiconvex;*

2.2 *m admits a filtering family, namely for every $B \in \Sigma$ there exists a filtering family $\{B_t\}_{t\in[0,1]}$ such that*

a) *$B_0 = \emptyset,\ B_1 = B$ and, if $t < t'$, then $B_t \subset B_{t'}$;*

b) *$m(B_t) = tm(B)$, for every $t \in [0,1]$;*

2.3 *m is a convex measure.*

Proof: 2.1) $\Longrightarrow$ 2.2). Let $B \in \Sigma$ be fixed. With a standard argument it is possible to construct a filtering sequence $(B_t)_t, t \in \mathbb{Q}(2)$ which satisfies conditions a) and b), see for example [6, Lemma 2.1]. Let now $t \in]0,1[$ be fixed, with $t \notin \mathbb{Q}(2)$, and let $(p_n)_n, (q_n)_n$ be two sequences in $\mathbb{Q}(2)$ such that $p_n \uparrow t$ and $q_n \downarrow t$. Put $B'_t = \cup_n B_{p_n},\ B''_t = \cap_n B_{q_n}$ and note that $B'_t \subseteq B''_t$; hence

$$m(B'_t) \geq \sup_n m(B_{p_n}) = tm(B) = \inf_n m(B_{q_n}) \geq m(B''_t) \geq m(B'_t).$$

Therefore we can choose for instance $B_t = \cap_n B_{q_n}$.

2.2) $\Longrightarrow$ 2.3). Let $A, B \in \Sigma$ be fixed and let $t \in [0,1]$. As in [7, Theorem 2.4] let

$$C_t = (B \setminus A)_t \cup (A \cap B) \cup (A \setminus B)_{1-t},$$

where the families $\{(B \setminus A)_t\}_t, \{(A \setminus B)_t\}_t$ are the filtering families for $B \setminus A$ and $A \setminus B$ respectively. We have that $C_0 = A, C_1 = B$ and

$$\begin{aligned} m(C_t) &= tm(B \setminus A) + m(A \cap B) + (1-t)m(B \setminus A) = \\ &= tm(B) + (1-t)m(A) \end{aligned}$$

2.3) $\Longrightarrow$ 2.1). It is obvious. □

The equivalence between 2.1) and 2.2) in the finite dimensional case had already been obtained in [7].

We shall suppose now that m is the indefinite integral of a function $e \in L^1_\mu(X^+)$ which satisfies the following assumption:

(h) μ is a *control* for m.

Remark 2 A sufficient condition for **(h)** is, for example, the following:

$$\operatorname{ess\,inf}_{\Omega} \|e\| \geq r > 0 \tag{3}$$

The assumption (3) has a very meaningful interpretation in the application to the economic model namely, when e is the initial endowment. We could label it as *minimal entrance feee* (m.e.f.) because from the economic point of view this means that all the consumers, except a set of measure zero, have a minimal granted support r.

Observe also that m.e.f. assumption is stronger than the condition **(h)**. For instance, consider in the interval $[0,1]$, $\mu : \Sigma \to [0,1]$ defined as

$$\mu(A) = \int_A \frac{1}{2\sqrt{x}} dx,$$

where dx is the Lebesgue measure and A is a Lebesgue measurable set and take $e(x) = 2\sqrt{x}$.
The measure $m = \int e d\mu$ coincides with the Lebesgue measure and admits μ as a control, but e does not satisfy m.e.f. assumption.

We want now to prove the announced result:

Proposition 3 *Let μ be a s-bounded, non negative, finitely additive measure, and let $e \in L^1_\mu(X^+)$ be such that $m = \int e d\mu$ is semiconvex and μ is a control for m; then the pairs (m, μ) and $(m, x\mu)$ are convex for every $x \in X$.*

Proof: The idea of the proof is analogous to that given in [7, Propositions 2.4 and 2.5]. First of all we prove that the pair $(m, |m|)$ is semiconvex.
Let $E \in \Sigma$ be fixed and consider the filtering family $\{E_t\}_{t\in[0,1]}$ associated to m and E. If $|m|(E_{1/2}) = 2^{-1}|m|(E)$ then we have semiconvexity. Otherwise let, for example, $|m|(E_{1/2}) < 2^{-1}|m|(E)$ and therefore $|m|(E \setminus E_{1/2}) > 2^{-1}|m|(E)$. If we apply Proposition 2 to m we find the filtering families $\{A_t\}_t, \{B_t\}_t$ associated to $E_{1/2}$ and $E \setminus E_{1/2}$ respectively. We set

$$C_t = A_t \cup B_{1-t}.$$

By construction, for every t

$$m(C_t) = tm(E_{1/2}) + (1-t)m(E \setminus E_{1/2}) = \frac{1}{2}m(E).$$

We prove now that $|m|(C_t)$ is a continuous function in t. Let y^*m be a Rybakov control for m as in Proposition 1. Therefore, for $\varepsilon > 0$ fixed, let $\delta > 0$ be that of the absolute continuity of $|m|$ with respect to y^*m. Let $t, s \in [0,1]$ be such that $|t-s|\ y^*m(E) \le \delta$; suppose for instance $t < s$. The set $C_t \ \Delta\ C_s$ is given by $(A_s \setminus A_t) \cup (B_{1-t} \setminus B_{1-s})$. Since $y^*m(C_t \ \Delta\ C_s) = |t-s|y^*m(E) < \delta$ we have that

$$|\ |m|(C_t) - |m|(C_s)\ | \ \le \ |m|(C_t \ \Delta\ C_s) < \varepsilon.$$

Hence, by continuity, as $|m|(C_0) < 2^{-1}|m|(E)$, while $|m|(C_1) > 2^{-1}|m|(E)$, there exists $t \in]0,1[$ such that $|m|(C_t) = 2^{-1}|m|(E)$.
Since μ is a control for m this implies also the continuity of $t \mapsto \mu(C_t)$ and then the semiconvexity of (m, μ). Finally the convexity of the pair (m, μ) follows by Proposition 2 if we consider $E = X \times I\!R$, $E^+ = X^+ \times [0, \infty[$, and the fact that the finitely additive measure (m, μ) is s-bounded. The convexity of $(m, x\mu)$ is an immediate consequence of the definition and of the convexity of (m, μ).

4. Applications to multivalued finitely additive integral of non convex integrands

Throughout this section, and similarly as in [23], we will adopt the following notations:

(i) (Ω, Σ) is a measurable space and $\mu : \Sigma \to [0, \infty[$ a s-bounded finitely additive measure which is also strongly continuous.

(ii) $(cf(X), h)$ and $(cwk(X), h)$ are the families of non empty, convex, closed (non empty, convex and weakly compact respectively) subsets of X with the Hausdorff distance.

(iii) $\Gamma = \sum_{i=1}^{p} C_i 1_{E_i}$ is a simple multifunction with closed and convex values with $E_i \cap E_j = \emptyset$ for $i \neq j$;

(iv) $e \in L^1_\mu(X^+)$ is such that $\lambda(E) := \int_E e d\mu$ is a convex finitely additive measure;

(v) $G = (\Gamma - e),\ F = G \cup \{0\}$.

We examine now the problem studied in [23] when the measure with respect to which we integrate is only finitely additive. This case is not a mere extension of the countably additive one. It has applications for instance in finitely additive economies which were introduced first in [1] by Armstrong and Richter: they explained why their model of a large economy is more realistic than the countably additive one. The same model was also extensively studied in [3], [4].

We begin with the bounded case, namely we assume that $C_i \in cwk(X)$ for every $i = 1, \ldots, p$. Thanks to the Radström Embedding Theorem ([24]) and to the Stone isomorphism we can consider the multifunctions $\overline{\Gamma}$ and $\overline{G}$ where

$$\overline{\Gamma} = \sum_{i=1}^{p} C_i 1_{h(E_i)}, \tag{4}$$

$$\overline{G} = \overline{\Gamma - e} = \sum_{i=1}^{p} C_i 1_{h(E_i)} - \overline{e}. \tag{5}$$

Thanks to (v) above, F certainly admits totally measurable selections and so we can define the Aumann integral as usual, using the finitely additive μ-integrability, [11, Chapter 3]: namely

- S^1_F is the set of all μ-integrable selections of a multifunction F, that is

$$S^1_F = \{f \in L^1_\mu(X) : f(\omega) \in F(\omega) \quad \mu - \text{almost everywhere}\};$$

- the Aumann integral of F is defined by

$$(A) - \int_E F d\mu = \left\{ \int_E f d\mu,\ f \in S^1_F \right\}.$$

The finitely additive case here considered is quite different from the case considered, for instance, in [21] and [22], where the existence of suitable selections depends upon the topological properties of the values of F.

Moreover, we shall denote by M_Γ, M_G the finitely additive multimeasures defined as the indefinite Aumann integrals of Γ and G respectively.
By $R_X(M_{(\cdot)})$ we shall denote the range of $M_{(\cdot)}$ that is

$$R_X(M_{(\cdot)}) = \bigcup_{E \in \Sigma} M_{(\cdot)}(E).$$

As Γ takes values in $cwk(X)$, its Aumann integral is convex and weakly compact: in fact, for every $E \in \Sigma$, we have that

$$\begin{aligned} \sum_{i=1}^{p} C_i \mu(E \cap E_i) \quad &\subseteq \quad (A) - \int_E \Gamma d\mu \subseteq (A) - \int_{h(E)} \overline{\Gamma} d\overline{\mu} = \\ &= \quad \sum_{i=1}^{p} C_i \overline{\mu}(h(E \cap E_i)), \end{aligned} \tag{6}$$

where the first inclusion and the last equality can be obtained by [9, Corollary 8] since in the proof of this corollary the countable additivity is not required, while the middle inclusion can be obtained analogously to the proof of [21, Theorem 5.1] (note that, since Γ is simple, we do not need the compactness of the values of Γ).

Using these facts, statements analogous to those of [23, Propositions 3.1, 3.2 and 3.4], with the strong continuity of μ replacing the non atomicity, hold for Γ also in the finitely additive case. Namely

(a) $M_\Gamma(E) = \left\{ \int_E s d\mu,\ s \text{ is a simple selection of } \Gamma \right\}$;

(b) $R_X(M_\Gamma) \in cwk(X)$;

(c) $M_G(E) = (A) - \int_E \Gamma d\mu - \int_E e d\mu \in cwk(X)$ for every $E \in \Sigma$.

In [23], Theorem 3.7 allowed us to obtain that $R_X(M_G) \in cwk(X)$ when μ is non atomic and countably additive. In the finitely additive case, Theorem 3.7 which is, in fact, a Lyapounov-type statement, does not hold in its complete extension (a counterexample is easily derived from [8, Theorem 4.4]). Its weakened finitely additive version, that is Theorem 1 below, will play a similar role in achieving the convexity of $R_X(M_G)$ in our case, although weak compactness is not assured in general. These two results, despite their similarity, do not compare: here we have only finite additivity but we have to assume the equivalence between μ and m; the proof of Theorem 1 completely differs from that of the quoted countably additive version, and is heavily based upon the results of Section 3.

Theorem 1 *Let X be a Banach lattice, Y a Banach space, μ a strongly continuous finitely additive measure and $e \in L^1_\mu(X^+)$ be such that the finitely additive measure $\lambda = \int e d\mu$ is convex and admits μ as a control. If f is a Y-valued simple function then setting $m = \int f d\mu$, the range of the pair (m, λ) is convex in $Y \times X$.*

Proof: We know that if $f = \sum_{i=1}^p c_i 1_{E_i}$ for some finite decomposition of Ω, $\{E_1, \ldots, E_p\}$ then

$$(m, \lambda)(\Sigma_E) = (m, \lambda)(\Sigma_{E \cap E_1}) + \ldots + (m, \lambda)(\Sigma_{E \cap E_p}).$$

So it is enough to note that, from Proposition 3, for each $i = 1, \ldots p$, $(m, \lambda)(\Sigma_{E \cap E_i})$ is convex and m is a multiple of μ on $\Sigma_{E \cap E_i}$. □

As an immediate consequence we obtain that

Corollary 1 *Let $f_j = \sum_{i=1}^p z_j^i 1_{E_i}, j = 1, 2$ be simple and measurable functions with values in X and let λ as in Theorem 1. Then setting $m_j = \int f_j d\mu - \lambda$ we have that $m_1, m_2, (m_1, m_2)$ are convex finitely additive measures.*

Proof: Applying Theorem 1, one easily deduces that m_1 and m_2 are convex finitely additive measures. We shall prove that the finitely additive measure (m_1, m_2) is convex. It is enough to prove that from Proposition 3, for each $i = 1, \dots p$, $(m_1, m_2)(\Sigma_{E\cap E_i})$ is convex for every $E \in \Sigma$.
By Proposition 3 the pair (λ, μ) is convex. So, for every $D_1, D_2 \in \Sigma_{E\cap E_i}$ and for every $t \in]0,1[$ there exists $C \in \Sigma_{E\cap E_i}$ such that

$$\begin{aligned} \mu(C) &= t\mu(D_1) + (1-t)\mu(D_2) \\ \lambda(C) &= t\lambda(D_1) + (1-t)\lambda(D_2). \end{aligned}$$

Since on $\Sigma_{E\cap E_i}$ $m_j = z_j^i \mu$ we have, for $j = 1, 2$,

$$\begin{aligned} m_j^i(C) &= z_j^i\mu(C) - \lambda(C) = \\ &= tz_j^i\mu(D_1) + (1-t)z_j^i\mu(D_2) - t\lambda(D_1) - (1-t)\lambda(D_2) = \\ &= tm_j(D_1) + (1-t)m_j(D_2); \end{aligned}$$

thus

$$(m_1, m_2)(C) = t(m_1, m_2)(D_1) + (1-t)(m_1, m_2)(D_2). \qquad \square$$

Because of Theorem 1 the set $R_X(M_G)$ is convex. In fact

Theorem 2 *Under the previous assumptions $R_X(M_G)$ is convex.*

Proof: It is possible to prove it in an analogous way as in [16, Lemma 7]. We report the proof for completeness. We recall that $G = \Gamma - e$, with $\Gamma = \sum_{i=1}^p C_i 1_{E_i}$ and $\lambda = \int e d\mu$.
Let $\varphi = \sum_{i=1}^p z_i 1_{E_i}$ be a fixed simple selection of Γ. Consider the finitely additive selection measure ν of M_G, defined as

$$\nu(\cdot) = \int_\cdot \varphi d\mu - \lambda.$$

Fix x_1, x_2 in $R_X(M_G)$ and $t \in]0,1[$. Then there exist two sets $A_1, A_2 \in \Sigma$ and two simple selections of Γ, f_1 and f_2 ($f_j = \sum_{i=1}^p z_j^i 1_{E_i}, j = 1, 2$), such that

$$x_j = \int_{A_j} f_j d\mu - \lambda(A_j), \qquad j = 1, 2.$$

We put $m_j = \int f_j d\mu - \lambda$, for $j = 1, 2$. By Corollary 1 $m_1, m_2, (m_1, m_2)$ are convex. Then there exist B_1, B_2, B_3 such that $B_1 \subset A_1 \setminus A_2$, $B_2 \subset A_2 \setminus A_1$ and $B_3 \subset A_1 \cap A_2$ and

$$m_1(B_1) = tm_1(A_1 \setminus A_2); \qquad m_2(B_2) = (1-t)m_2(A_2 \setminus A_1);$$
$$(m_1, m_2)(B_3) = t(m_1, m_2)(A_1 \cap A_2).$$

Set $B_4 = (A_1 \cap A_2) \setminus B_3$ and $B = \cup_{i=1}^4 B_i$. We will show that for a suitable selection ν^* of M_G, $\nu^*(B) = tx_1 + (1-t)x_2$.
Indeed, for every $E \in \Sigma$, define

$$\begin{aligned}
\nu^*(E) &= \nu(E \setminus B) + m_1(E \cap (B_1 \cup B_3)) + m_2(E \cap (B_2 \cup B_4)) = \\
&= \int_{E \setminus B} \varphi d\mu - \lambda(E \setminus B) + \int_{E \cap (B_1 \cup B_3)} f_1 d\mu + \\
&+ \int_{E \cap (B_2 \cup B_4)} f_2 d\mu - \lambda(E \cap B) = \\
&= \int_E \left(\varphi \cdot 1_{E \setminus B} + f_1 \cdot 1_{E \cap (B_1 \cup B_3)} + f_2 \cdot 1_{E \cap (B_2 \cup B_4)}\right) d\mu - \lambda(E).
\end{aligned}$$

Then the finitely additive measure ν^* is a selection of M_G and if we evaluate ν^* on the set B we obtain

$$\begin{aligned}
\nu^*(B) &= m_1(B_1 \cup B_3) + m_2(B_2 \cup B_4) = tm_1(A_1 \setminus A_2) + tm_1(A_1 \cap A_2) + \\
&+ (1-t)m_2(A_2 \setminus A_1) + m_2(A_1 \cap A_2 \setminus B_3) = \\
&= tx_1 + (1-t)x_2. \qquad \square
\end{aligned}$$

Moreover we have that

Theorem 3 *If G is a multifunction as before then $cl\{R_X(M_G)\} = R_X(M_{\overline{G}})$.*

Proof: From (c), (5), (6) above and [22, Theorem 5.1], (which holds in our case without compactness of values of G), for every $E \in \Sigma, M_G(E) \subset M_{\overline{G}}(\tau(E))$. For the converse inclusion, if $x \in R_X(M_{\overline{G}})$ then there exists a set $H \in \mathcal{G}_\sigma$ such that

$$x \in M_{\overline{G}}(H) = (A) - \int_H \overline{\Gamma} d\overline{\mu} - \overline{\lambda}\ (H)$$

and then $x = \sum_{i=1}^{p} x_i \overline{\mu}(H \cap \tau(E_i)) - \overline{\lambda}\ (H)$ for some $x_i \in C_i,\ i = 1, \ldots, p$.
Since $\mathcal{G}$ is Fréchet-Nikodym dense in $\mathcal{G}_\sigma$, for every $\varepsilon > 0$ there exists a $\mathcal{G}$ -measurable set B such that

$$\overline{\mu}(H \Delta B) \leq \frac{\varepsilon}{2kp},$$

for some $k > \max\{\|e\|_1, h(C_1, \{0\}), \ldots, h(C_p, \{0\})\}$. Then

$$\|x - \sum_{i=1}^{p} x_i \overline{\mu}(B \cap \tau(E_i)) - \overline{\lambda}\ (B)\| \leq \sum_{i=1}^{p} \|x_i\| \overline{\mu}(H \Delta B) + |\overline{\lambda}\ |(H \Delta B) \leq \varepsilon.$$

Since B is a $\mathcal{G}$ -measurable set, there exists $E \in \Sigma$ such that $\tau(E) = B$; let us put

$$x_\varepsilon := \sum_{i=1}^{p} x_i \overline{\mu}(B \cap \tau(E_i)) - \overline{\lambda}\ (B) = \sum_{i=1}^{p} x_i \mu(E \cap E_i) - \lambda(E).$$

We have that $x_\varepsilon \in R_X(M_G)$ and $\|x - x_\varepsilon\| \leq \varepsilon$.
Hence $R_X(M_{\overline{G}}) \subset cl\{R_X(M_G)\}$ and again applying (c), the equality follows. □

Now, as in [23, Theorem 3.15], one can show that, if $0 \notin G(\omega)$ for all $\omega \in \Omega$, then, for every $E \in \Sigma$,

$$(A) - \int_E F d\mu = R_X(M_G|_{\Sigma \cap E}) \tag{7}$$

and therefore it is convex.

Let us now consider the unbounded case, namely assume that $C_i \in cf(X),\ i = 1, \ldots p$. For each integer n, consider

$$\Gamma_n(\omega) = \Gamma(\omega) \cap nX_1, \qquad F_n(\omega) = (\Gamma_n(\omega) - e(\omega)) \cup \{0\}.$$

As in [23, Proposition 3.16] one shows that, for every $E \in \Sigma$

$$(A) - \int_E F d\mu = \bigcup_n\ (A) - \int_E F_n d\mu. \tag{8}$$

The sequence on the right hand side of (8) is increasing. Hence, if $0 \notin G$, $(A) - \int_E F d\mu$ is the union of an increasing sequence of convex sets, and therefore is convex.

For the general case, that is when some values of G contain 0, similarly to [23, Theorem 3.18] denote by Ω_0 the set $\{\omega : 0 \in G(\omega)\}$; the Aumann integral splits into two parts:

$$(A) - \int_E F d\mu = (A) - \int_{E\cap\Omega_0} F d\mu + (A) - \int_{E\setminus\Omega_0} F d\mu$$

It only remains to derive the convexity of the first summand by noting that F has convex values in Ω_0.

In conclusion we have obtained the following:

Theorem 4 *Let (Ω, Σ, μ) be a non negative finitely additive measure space with μ strongly continuous, X a Banach lattice, $\Gamma : \Omega \to cf(X)$ a simple multifunction and $e \in L^1_\mu(X^+)$ generate a semiconvex finitely additive measure which admits μ as a control. Consider $F = (\Gamma - e) \cup \{0\}$. Then, for every $E \in \Sigma$, $(A) - \int_E F d\mu$ is convex.*

We shall now derive a result similar to [23, Theorem 3.18]. Indeed, by means of Theorem 3, and the results in the countably additive case, [23, Theorem 3.18], we obtain the following

Theorem 5 *Let (Ω, Σ, μ) be a non negative finitely additive measure space with μ strongly continuous, X a Banach lattice, $\Gamma : \Omega \to cf(X)$ a simple multifunction and let $e \in L^1_\mu(X^+)$ generate a Liapounov indefinite integral which admits μ as a control. Consider $F = (\Gamma - e) \cup \{0\}$. Then, for every $E \in \Sigma$, $(A) - \int_E F d\mu$ is a convex set which is the union of an increasing sequence of convex and relatively weakly compact sets.*

Proof: The line of the proof is somewhat analogous to that of [23, Theorem 3.18]. As before consider

$$\Gamma_n(\omega) = \Gamma(\omega) \cap nX_1, \quad G_n(\omega) = \Gamma_n(\omega) - e(\omega), \quad F_n = G_n(\omega) \cup \{0\}.$$

and the Stone transforms of Γ_n, G_n, denoted by $\overline{\Gamma}_n$ and $\overline{G}_n$ respectively. From [23, Theorem 3.14] $R_X(M_{\overline{G}_n})$ is weakly compact and, since from (7),

$$(A) - \int_E F_n d\mu = R_X(M_{G_n|_{E\cap\Sigma}})$$

by Theorem 3 above, we have that $cl\left\{\int_E F_n d\mu\right\}$ is convex and weakly compact for each n. Hence the already proved equality

$$(A)-\int_E F d\mu = \bigcup_n (A)-\int_E F_n d\mu$$

shows the assertion. □

Remark 3 Note that Theorem 5 above is the only result that can be partially derived from the countably additive case [23] by means of the Stone extension. The other results in this section, despite the similarity of the statements, cannot be obtained in this way since the assumptions here and in [23] do not compare. For example if we take μ the Lebesgue measure on $[0,1]$, a non negative integrable function e such that $\mu(\text{supp } e) \in]0,1[$ then $\lambda = \int e d\mu$ is automatically Liapounov but $\mu \not\ll \lambda$. Viceversa the example which can be derived from [8] verifies the hypothesis on μ and λ given here but λ is not Liapounov.

Acknowledgment

This work was partially supported by the G.N.A.M.P.A. of I.N.D.A.M.
The authors would like to thank one of the referees for his suggestions and improvements of some of the results obtained in this paper.

References

[1] T.E. ARMSTRONG–M.K. RICHTER, The Core-Walras equivalence, *J. Econom. Theory,* **33**, 116-151, (1984).

[2] A. AVALLONE–A. BASILE, On the Liapunov–Richter theorem in finitely additive setting, *J. of Mathematical Economics,* **22**, 557-561, (1993).

[3] A. BASILE, Finitely additive correspondences, *Proc. of the Amer. Math. Soc.,* **121**, N. 3, 883-891, (1994).

[4] A. BASILE–M.G. GRAZIANO, Restricted coalition formation mechanism in finitely additive economies, preprint 2001.

[5] C. L. BYRNE, Remarks on the Set-Valued Integrals of Debreu and Aumann, *J. Math. Anal. Appl.,* **62**, 243-246, (1978).

[6] D. CANDELORO–A. MARTELLOTTI, Su alcuni problemi relativi a misure scalari subadditive e applicazioni al caso dell'additività finita, *Atti Sem. Mat. Fis. Univ. Modena,* **27**, 284-296, (1978).

[7] D. CANDELORO–A. MARTELLOTTI, Sul rango di una massa vettoriale, *Atti Sem. Mat. Fis. Univ. Modena,* **28**, 102-111,(1979).

[8] D. CANDELORO–A. MARTELLOTTI, Geometric Properties of the Range of Two-Dimensional Quasi-Measures with respect to Radon-Nikodym Property, *Adv. in Math.,***93** N.1, 9-24,(1992).

[9] J. DIESTEL–J.J. UHL, *Vector measures,* Math. Surveys **15**, Amer. Math. Soc., Providence, 1977.

[10] L. DREWNOWSKI, On control submeasures and measures, *Studia Math.,* **50**, 203-224, (1974).

[11] N. DUNFORD–J. T. SCHWARTZ, *Linear Operators Part I,* Interscience, New York, 1958.

[12] C. FEFFERMANN, L_p spaces over finitely additive measures, *Pacific J. Math.,* **26**, 265-271, (1968).

[13] W. HILDEBRAND, *Core and Equilibria of a Large Economy,* Princeton University Press, Princeton, 1974.

[14] M. P. KATS, On the continuation of vector measures, *Sibirskii Mathematicheskii Zhurnal,* **13** N. 5, 1158-1168, (1972).

[15] M.C. ISIDORI–A. MARTELLOTTI–A.R. SAMBUCINI, Integration with respect to orthogonally scattered measures, *Math. Slovaca,* **48** N. 3, 253-269, (1998).

[16] LE VAN TU, The range of a multimeasure, *J. Australian Math. Soc.,* (Serie A) **26**, 442-452, (1978).

[17] J. LINDENSTRAUSS, A short proof of Liapounoff's convexity theorem, *J. of Mathematics and Mechanics,* **6** , 971–972, (1966).

[18] D. MAHARAM, Finitely additive measures on integers, *Indian J. of Statistic,* **38** Serie A, Pt.1, 44-59, (1976).

[19] A. MARTELLOTTI, Finitely additive Phenomena, *Rend. Ist. Mat. Univ. Trieste,* **33**, 201-249, (2001).

[20] A. MARTELLOTTI–K. MUSIAL–A.R. SAMBUCINI, A Radon-Nikodym theorem for the Bartle-Dunford-Schwartz Integral with respect to Finitely Additive Measures, *Atti Sem. Mat. Fis. Univ. Modena,* **42**, 625-633, (1994).

[21] A. MARTELLOTTI–A.R. SAMBUCINI, On the comparison between Aumann and Bochner integrals, *J. of Math. Anal. and Appl.,* **260**, 6-17, (2001).

[22] A. MARTELLOTTI–A.R. SAMBUCINI, The finitely additive integral of multifunction with closed and convex values, *Zeitschrift für Analysis ihre Anwendungen,* **21**, 851-864, (2002).

[23] A. MARTELLOTTI–A.R. SAMBUCINI, Multivalued integral of non convex integrands, *Internation J. of Pure and Applied Mathematics,* **5** N. 1, 11-28, (2003).

[24] H. RADSTRÖM, An Embedding Theorem for Spaces of Convex Set, *Proc. Amer. Math. Soc.* **3** , 165–169, (1952).

[25] V. RYBAKOV Theorem of Bartle, Dunford and Schwartz on vector valued measures, *Math. Notes,* **7***, 147-151, (1970).*

[26] A. R. SAMBUCINI, Remarks on set valued integrals of multifunctions with non empty, bounded, closed and convex values, *Comm. Math.,* **39**,153-165, (1999).

[27] R. SIKORSKI, *Boolean Algebras,* Springer Verlag, Berlin, 1964.

[28] K. VIND, Edgeworth Allocations in Exchange Economy with many Traders, *Int. Econom. Review,* **5** , 165-177, (1964).

[29] H. VOLKMER–H. WEBER, Der Wertebereich atomloser Inhalte, *Archive der Mathematik,* **40**, 464-474, (1983).

Instructions to Contributors
Journal of Concrete and Applicable Mathematics

A quartely international publication of Eudoxus Press, LLC, of TN.

Editor in Chief: George Anastassiou
Department of Mathematical Sciences
University of Memphis
Memphis, TN 38152-3240, U.S.A.

1. Manuscripts hard copies in triplicate, and in English, should be submitted to the Editor-in-Chief:

Prof.George A. Anastassiou
Department of Mathematical Sciences
The University of Memphis
Memphis,TN 38152, USA.
Tel.001. 901.678.3144
e-mail:ganastss@memphis.edu.

Authors may want to recommend an associate editor the most related to the submission to possibly handle it.
Also authors may want to submit a list of six possible referees, to be used in case we cannot find related referees by ourselves.
2. Manuscripts should be typed using any of TEX,LaTEX,AMS-TEX,or AMS-LaTEX and according to EUDOXUS PRESS, LLC. LATEX STYLE FILE. (VISIT www.msci.memphis.edu/~ganastss/jcaam / to save a copy of the style file.)They should be carefully prepared in all respects. Submitted copies should be brightly printed (not dot-matrix), double spaced, in ten point type size, on one side high quality paper 8(1/2)x11 inch. Manuscripts should have generous margins on all sides and should not exceed 24 pages.
3. Submission is a representation that the manuscript has not been published previously in this or any other similar form and is not currently under consideration for publication elsewhere. A statement transferring from the authors(or their employers,if they hold the copyright) to Eudoxus Press, LLC, will be required before the manuscript can be accepted for publication.The Editor-in-Chief will supply the necessary forms for this transfer.Such a written transfer of copyright,which previously was assumed to be implicit in the act of submitting a manuscript,is necessary under the U.S.Copyright Law in order for the publisher to carry through the dissemination of research results and reviews as widely and effective as possible.
4. The paper starts with the title of the article, author's name(s) (no titles or degrees), author's affiliation(s) and e-mail addresses. The affiliation should comprise the department, institution (usually university or company), city, state (and/or nation) and mail code.
The following items, 5 and 6, should be on page no. 1 of the paper.

5. An abstract is to be provided, preferably no longer than 150 words.
6. A list of 5 key words is to be provided directly below the abstract. Key words should express the precise content of the manuscript, as they are used for indexing purposes. The main body of the paper should begin on page no. 1, if possible.
7. All sections should be numbered with Arabic numerals (such as: 1. INTRODUCTION) .
Subsections should be identified with section and subsection numbers (such as 6.1. Second-Value Subheading).
If applicable, an independent single-number system (one for each category) should be used to label all theorems, lemmas, propositions, corrolaries, definitions, remarks, examples, etc. The label (such as Lemma 7) should be typed with paragraph indentation, followed by a period and the lemma itself.
8. Mathematical notation must be typeset. Equations should be numbered consecutively with Arabic numerals in parentheses placed flush right,and should be thusly referred to in the text [such as Eqs.(2) and (5)]. The running title must be placed at the top of even numbered pages and the first author's name, et al., must be placed at the top of the odd numbed pages.
9. Illustrations (photographs, drawings, diagrams, and charts) are to be numbered in one consecutive series of Arabic numerals. The captions for illustrations should be typed double space. All illustrations, charts, tables, etc., must be embedded in the body of the manuscript in proper, final, print position. In particular, manuscript, source, and PDF file version must be at camera ready stage for publication or they cannot be considered.
Tables are to be numbered (with Roman numerals) and referred to by number in the text. Center the title above the table, and type explanatory footnotes (indicated by superscript lowercase letters) below the table.
10. List references alphabetically at the end of the paper and number them consecutively. Each must be cited in the text by the appropriate Arabic numeral in square brackets on the baseline.

References should include (in the following order):
initials of first and middle name, last name of author(s)
title of article,
name of publication, volume number, inclusive pages, and year of publication.
Authors should follow these examples:

Journal Article

1. H.H.Gonska,Degree of simultaneous approximation of bivariate functions by Gordon operators, (journal name in italics) *J. Approx. Theory*, 62,170-191(1990).

Book

2. G.G.Lorentz, (title of book in italics) *Bernstein Polynomials* (2nd ed.), Chelsea,New York,1986.

Contribution to a Book

3. M.K.Khan, Approximation properties of beta operators,in(title of book in italics) *Progress in Approximation Theory* (P.Nevai and A.Pinkus,eds.), Academic Press, New York,1991,pp.483-495.

11. All acknowledgements (including those for a grant and financial support) should occur in one paragraph that directly precedes the References section.
12. Footnotes should be avoided. When their use is absolutely necessary, footnotes should be numbered consecutively using Arabic numerals and should be typed at the bottom of the page to which they refer. Place a line above the footnote, so that it is set

off from the text. Use the appropriate superscript numeral for citation in the text.
13. After each revision is made please again submit three hard copies of the revised manuscript, including in the final one. And after a manuscript has been accepted for publication and with all revisions incorporated, manuscripts, including the TEX/LaTex source file and the PDF file, are to be submitted to the Editor's Office on a personal-computer disk, 3.5 inch size. Label the disk with clearly written identifying information and properly ship, such as:

Your name, title of article, kind of computer used, kind of software and version number, disk format and files names of article, as well as abbreviated journal name.

Package the disk in a disk mailer or protective cardboard. Make sure contents of disks are identical with the ones of final hard copies submitted!

Note: The Editor's Office cannot accept the disk without the accompanying matching hard copies of manuscript. No e-mail final submissions are allowed! The disk submission must be used.

14. Effective 1 Nov. 2005 the journal's page charges are $8.00 per PDF file page. Upon acceptance of the paper an invoice will be sent to the contact author. The fee payment will be due one month from the invoice date. The article will proceed to publication only after the fee is paid. The charges are to be sent, by money order or certified check, in US dollars, payable to Eudoxus Press, LLC, to the address shown in the Scope and Prices section.

No galleys will be sent and the contact author will receive one(1) electronic copy of the journal issue in which the article appears.

15. This journal will consider for publication only papers that contain proofs for their listed results.

TABLE OF CONTENTS,JOURNAL OF CONCRETE AND APPLICABLE MATHEMATICS,VOL.3,NO.4,2005

WEIGHTED INTEGRAL INEQUALITIES IN TWO DIMENSIONS,
G.HANNA,J.ROUMELIOTIS,……..389

ON A CLASS OF POLYNOMIALS GENERALIZING THE LAGUERRE FAMILY,G.DATTOLI,P.RICCI,C.CESARANO,…….405

ASYMPTOTIC FORMULAE FOR POSITIVE LINEAR OPERATORS ON CONVEX SUBSETS OF BANACH SPACES,
F.ALTOMARE,S.DIOMEDE,……413

GENERALISED WEIGHTED TRAPEZOIDAL RULES AND RELATIONSHIP TO OSTROWSKI RESULTS,
P.CERONE.J.ROUMELIOTIS,……..437

PARALLEL SCHEMES FOR TWO-DIMENSIONAL PARABOLIC EQUATION WITH A BOUNDARY INTEGRAL CONDITION,M.DEHGHAN,…….457

RANDOM FIXED POINTS AND ITERATION PROCESS FOR ASYMPTOTICALLY NONEXPANSIVE RANDOM MAPS,I.BEG,……469

A NOTE ON A LIAPOUNOV-LIKE THEOREM FOR SOME FINITELY ADDITIVE MEASURES AND APPLICATIONS,
A.MARTELLOTTI,A.SAMBUCINI,……..481